Network reconstruction and causal analysis in systems biology

Séverine Affeldt

THÈSE DE DOCTORAT DE
l'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

Informatique

École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Séverine AFFELDT

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Reconstruction de Réseaux Fonctionnels et Analyse Causale en Biologie des Systèmes

(Network Reconstruction and Causal Analysis in Systems Biology)

soutenue le 02 juillet 2015

devant le jury composé de :

Dr. Florence Jaffrezic	Rapporteur
Pr. Philippe Leray	Rapporteur
Pr. Christophe Gonzales	Examinateur
Pr. Patrick E. Meyer	Examinateur
Dr. Loïc Paulevé	Examinateur
Dr. Hervé Isambert	Directeur de thèse

GJDX XTKQ PKXE KQDS RJTZ QOHZ OTW,
N JDCK OHZQ MZSM NWTY,
QQIX YSUZ IAIW MRLU OYFQ XRPB BEG[1]

Alan Turing

[1] Enigma machine settings

Ref	Wheels	Ring	Ground	Plugs
C	II VIII IV	XAQ	RFV	DV FE GP HO JW RM SZ TN UI XL

PIERRE AND MARIE CURIE UNIVERSITY

Abstract

Network Reconstruction and Causal Analysis in Systems Biology

by Séverine AFFELDT

The inference of causality is an everyday life question that spans a broad range of domains for which interventions or time-series acquisition may turn out to be impracticable if not unethical. Yet, elucidating causal relationships within *real-life* complex systems can be rather convoluted when relying exclusively on observational data. In this dissertation, I report a novel network reconstruction method, which combines constraint-based and Bayesian frameworks to reliably reconstruct graphical models despite inherent sampling noise in finite observational datasets. The approach is based on an information theory result tracing back the existence of colliders in graphical models to negative conditional 3-point information between observed variables. In turn, this provides a confident assessment of structural independencies in causal graphs, based on the ranking of their most likely contributing nodes with (significantly) positive conditional 3-point information. Starting from a complete undirected graph, dispensible edges are progressively pruned by iteratively *taking off* the most likely positive conditional 3-point information from the 2-point (mutual) information between each pair of nodes. The resulting network skeleton is then partially directed by orienting and propagating edge directions, based on the sign and magnitude of the conditional 3-point information of unshielded triples. This 3off2 network reconstruction approach is shown to outperform constraint-based, search-and-score and earlier hybrid methods on a range of benchmark networks. In addition, the 3off2 method provides promising predictions when applied to the reconstruction of complex biological systems, such as hematopoietic regulatory subnetworks, zebrafish neural networks, mutational pathways driving tumour progression or the interplay of genomic properties on the evolution of vertebrates.

Acknowledgements

I would like to thank all the people who have contributed to turn this Ph.D. project into a unique human and scientific adventure.

First and foremost, I am deeply grateful to my supervisor, Hervé Isambert, whose unwavering support and trust have continuously guided me on the *somewhat* convoluted path of science. Thank you so much for your permanent availability and precious involvement in this ambitious and increasingly promising project that we have begun few years ago!

I would like to express my sincere gratitude to Florence Jaffrezic and Philippe Leray, for their precious time reading my manuscript and their constructive and expert comments. I would like to extend my gratitude to Christophe Gonzales for having given me the honour to chair my thesis committee. I also wish to thank Loïc Paulevé and Patrick Meyer for triggering interesting discussions.

I am also grateful to Sylvie Coscoy, Jacques Camonis and Martin Weigt for accepting to be my tutors during these four years of thesis. Thank you for your scientific insights and advice. I would also like to thank the professors from UPMC, Nizar Ouarti, Alessandra Carbone, Béatrice Bérard and Cécile Le Page, for offering me the exciting opportunity to teach during my Ph.D. project, and all the UPMC students of these courses for their steady dynamic involvement. I am also thankful to my former colleagues at Dassault Systèmes, and in particular Sylvain Huzé and Michel Latappy for their encouraging enthusiasm when I shared with them my plan of *getting back to school*.

It would have not been possible to pursue such a wide range of captivating bioinformatic projects without the help of many master trainees. Thank you Samya, Maria, Francesca, Jatin, Wenjun and Fazal for your invaluable contribution! I am also very grateful to Sasha, Atoussa, Sébastien and Mica for their friendly past and ongoing collaborations.

Workdays would have been far less inspiring without my dear colleagues and friends, especially Giulia and Param, with whom I have shared most of my stay at the Institute. Thanks a lot for the thrilling discussions and, most importantly, for participating in building such a nice team spirit. I would also like to thank Louis for his numerous biological inputs, and above all, for his kindness. Many thanks also go to the people with whom I had the chance to share my office, Démosthène, Nada, Grégory, Vasilica, Marzuk, Senthil, Simon, Aurore, Sabrina, Vasily, Daniel, Adrien and Guillaume. I would like to address many thanks to the efficient administrative team of the Institute, and in particular Agnès for being so generous with her time when it comes to helping others.

I would like to thank many friends outside the Institute for their everyday support and affection. Thanks a lot Yoriko, Fanny, Emilie, Clémentine and Julien for all these nice moments that repeatedly helped me to put any bad day into perspective, and especially Patrick for all these years of kind patience and support.

These acknowledgements would be incomplete without mentioning my family. I am infinitely indebted to my parents who showed me how important and rewarding it could be to focus energy and will on something you strongly believe in, and to my brother, who taught me that it inevitably takes a year to make a day.

Contents

To my parents

Chapter 1

Introduction

When I joined the Physical-Chemistry Curie Laboratory in 2011, the Isambert's research group was deeply involved in the understanding of a surprising observation. 'Dangerous' gene families, defined as prone to dominant (gain-of-function) mutations[1], have been greatly expanded in the course of vertebrate evolution by contrast to gene families not implicated in diseases or prone to recessive (loss-of-function) mutations[2]. From an evolutionary perspective, this expansion of 'dangerous' gene families, implicated in cancer and other severe genetic diseases in human, is undoubtedly puzzling as it enhances the susceptibility of vertebrates to dominant genetic diseases. A first insight into the underlying evolutionary process came with the realization that the expansion of dangerous gene families resulted from very particular and dramatic accidents whereby the whole genome of a species is duplicated at once. Two rounds of such whole genome duplication (WGD) occurred at the origin of vertebrates some 500MY ago [Ohno, 1970, Putnam et al., 2008]. But could gene susceptibility to dominant deleterious mutations be somehow *responsible* for the striking evolutionary retention of 'dangerous' gene families after WGD?

Dropped from a computational background, but truly motivated by issues related to evolution, I was immediately thrilled by this biological problem that highlights a surprising connection between two fairly different time scales. On the one hand, some genes, once mutated, play a crucial role in the development of complex diseases and hence, have a severe impact at the level of one individual's life span.

[1] Dominant mutations lead to constitutively active mutant genes with dominant phenotypes typically detrimental for the organism. In particular, the oncogenes are prone to dominant deleterious mutations.

[2] Recessive mutations are mutations that lead to a loss of function. In a diploid organism, both alleles need to be mutated in order to express detrimental phenotypes, as the (functional) non mutated allele can usually maintain the warrant functionality.

On the other hand, the reason why such genes remain in our genome, despite their potential harmfulness, could be found in the course of vertebrate evolution that begins hundreds of millions years ago.

The first part of my thesis was thus dedicated to the establishment of a novel evolutionary model suggesting that this somewhat counterintuitive expansion of 'dangerous' gene families is in fact a *consequence* of their susceptibility to deleterious mutations and purifying selection in polyploid species that arose from two rounds of whole genome duplication (WGD) events, dating back from the onset of jawed vertebrates, some 500MY ago [Ohno, 1970, Putnam et al., 2008]. In a nutshell, the evolutionary mechanism we have proposed relies on the fact that all the WGD-duplicates are initially acquired through speciation without the need to provide evolutionary benefit to be fixed in post-WGD species (section 1.1 & Chapter 8). This part of my research project, which has been the initial motivation for the second and main part of my PhD, has led to three publications (appended at the end of Chapter 1).

The main statistical approach I have used to analyse the retention of WGD-duplicates is the Mediation analysis, formalized by Pearl [Pearl, 2001, 2012]. It enabled me to disentangle and quantify the *direct* and *indirect* causal effects between the susceptibility to dominant deleterious mutations, the retention of WGD-duplicates, so-called 'ohnologs'[3], and a third genomic property chosen among the ones proposed so far to explain the retention of ohnologs (section 1.1 & Chapter 8). Yet, the Mediation analysis requires an oriented graphical model as input and does not directly uncover causal insights from observational data. Thus, I turned to the field of artificial intelligence and uncertainty in order to infer causal graphical models that could answer complex biological problems, such as the retention of ohnologs. In particular, beyond the susceptibility to dominant deleterious mutations, many other genomic properties, such as gene essentiality [4], functional ontology, expression level, divergence rates, etc, that are all correlated to some extent with the fate of gene duplicates throughout vertebrate evolution, needed also to be taken into account.

This led me to the second part of my research project, which is the bulk of my thesis, namely the reconstruction of reliable causal graphical models from finite

[3] In his seminal book, Susumu Ohno laid down the theoretical framework of the evolution after gene duplication [Ohno, 1970]. He proposed that genome duplications are a significant mechanism of evolution, even in the animal genomes [Ohno, 1970, Ohno et al., 1968]. The copies of genes arising from a whole genome duplication are now coined 'ohnologs', after his name.

[4] Essential genes are genes for which the silencing of the two alleles lead to lethality or sterility. The *essential* genes are typically prone to recessive mutations.

observational data, an utterly important issue in bioinformatics and other computational fields. In this part, I have developed a robust approach to reconstruct graphical models from finite datasets that combines the advantages of the state-of-the-art constraint-based (Chapter 3) and Bayesian (Chapter 4) methods. This new hybrid approach, named 3off2, is based on an information theoretic framework (Chapter 5) to confidently ascertain structural independencies in causal graphs and skeleton orientations (Chapter 6) with clear improvements over both constraint-based and Bayesian methods on benchmark networks (Chapter 7).

In the following section, I will briefly summarize the first part of my thesis (see reprints of papers at the end of this chapter for details) and also introduce one of the first applications for which the 3off2 method has brought valuable insights (Chapter 8). Yet, as shown later in this manuscript, the 3off2 method can be applied to a wide range of complex biological problems, such as the reconstruction of neural networks (Chapter 9), the discovery of transcription factors regulation networks (Chapter 10) or mutational pathways during tumour progression (Chapter 11). In the remaining of this chapter, I will introduce two main methods for causal graphical model reconstruction. Further details are provided in Chapters 3, 4 & 5. Finally, I will briefly outline how the 3off2 approach combines the state-of-the-art methods and provides a more robust scheme to reconstruct causal graphical models from finite observational datasets. The complete description of the 3off2 approach is detailed in Chapter 6.

1.1 Expansion of 'dangerous' gene families in vertebrates

Whole genome duplication (WGD) events have now been established in all major eukaryotes kingdoms. Two rounds of WGD in early vertebrates are frequently credited with creating the condition for the evolution of vertebrate complexity. In the first part of my thesis, I have shown that the two rounds of WGD in vertebrates have led to the preferential expansion of 'dangerous' gene families, defined as prone to dominant deleterious mutations. The details of this work are left in the reprints of the publications appended at the end of this Chapter. These results suggest that the striking expansion of gene families implicated in cancer and other severe genetic diseases is in fact a consequence of their susceptibility to deleterious mutations and the ensuing purifying selection in post-WGD species.

Our data mining analysis, based on the $20,506$ human protein coding genes, first revealed a strong correlation between the retention of WGD duplicates, and their

susceptibility to dominant deleterious mutations in human [Affeldt et al., 2013, Singh et al., 2012]. It appears that the human genes associated with the occurrence of cancer and other genetic diseases (8,095) have retained significantly more ohnologs than expected by chance (48% *versus* 35%; 48% : 3,844/8,095; $P = 1.3 \times 10^{-128}$, χ^2). I have also found that the retention of ohnolog is more strongly related to their 'dangerousness' than their 'essentiality' [Singh et al., 2012, 2014] (Chapter 8).

To go beyond mere correlations, I have then performed Mediation Analyses, following the approach of Pearl [Pearl, 2001, 2012], and quantified the *direct* and *indirect* effects of many genomic properties, such as essentiality, expression levels or divergence rates, on the retention of ohnologs. This enabled the investigation of an alternative hypothesis frequently invoked to account for the biased retention of ohnologs, namely the 'dosage-balance' hypothesis [Makino and McLysaght, 2010]. While this hypothesis posits that the ohnologs are retained because their interactions with protein partners require to maintain balanced expression levels throughout evolution, it appeared that the ohnologs have in fact been eliminated from permanent complexes in human (7.5% *versus* 35%; 7.5% : 18/239; $P = 1.2 \times 10^{-18}$, χ^2). The Mediation Analyses also showed (Fig. 1.1) that the gene susceptibility to deleterious mutations is more relevant than dosage-balance for the retention of ohnologs in more transient complexes [Singh et al., 2012] (Chapter 8).

These results suggest that the retention of human ohnologs is primarily caused by their susceptibility to deleterious mutations. They further establish that the retention of many ohnologs suspected to be dosage balanced is in fact *mediated* by their susceptibility to dominant deleterious mutations. All in all, this supports a new evolutionary model relying on a non-adaptive mechanism that hinges on (i) the *speciation* event concomitant to WGD, and (ii) the *dominance* of deleterious mutations leading to purifying selection in post-WGD species (see appended publications & Chapter 8). These results exemplify the role of non-adaptive forces on the emergence of eukaryotes complexity.

1.2 Learning causal graphical models

The prospect of learning the direction of causal dependencies from mere correlations in observational data has long defied practical implementations [Reichenbach, 1956]. The fact that causal relationships can, to some extent, be inferred from nontemporal statistical data is now known to hinge on the unique statistical

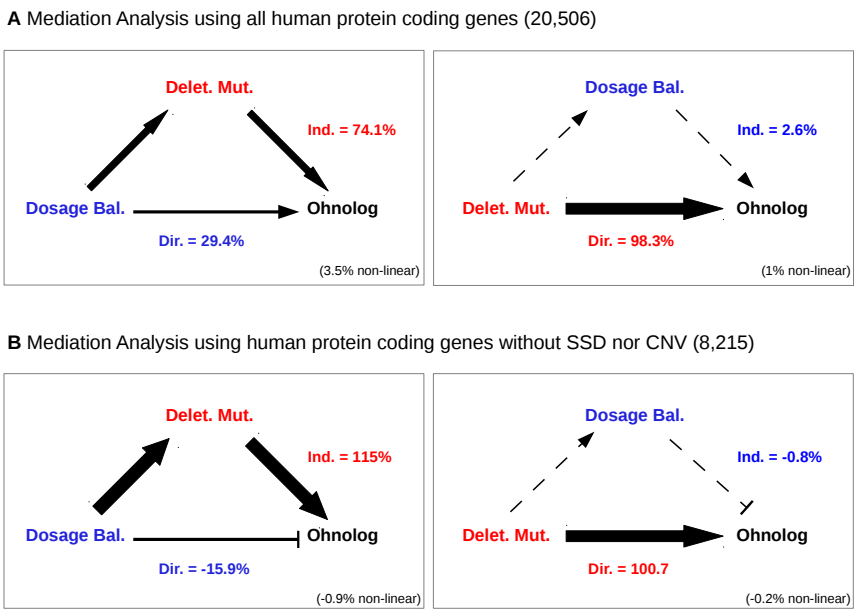


FIGURE 1.1 Quantitative Mediation analysis of direct versus indirect effects of deleterious mutations and dosage balance on the retention of human ohnologs using (A) all human protein coding genes (20,506) or (B) human protein coding genes without small scale duplicates (SSD) nor copy number variants (CNV) (8,215). The thickness of the arrows outlines the relative importance of the corresponding direct or indirect effects. $Dir. < 0$ or $Ind. < 0$ corresponds to an anticorrelated direct or indirect effect, respectively. Gene prone to deleterious mutations and/or to dosage balance (including haploinsufficient genes and genes involved in multiprotein complexes) are taken from [Singh et al., 2012].

imprint of colliders in causal graphical models, provided that certain assumptions are made about the underlying process of data generation, such as its faithfulness to a tree structure [Rebane and Pearl, 1988] or a directed acyclic graph model [Pearl, 2009, Spirtes et al., 2000] (see Chapter 2). These early findings led to the developments of two types of network reconstruction approaches that have been applied to a variety of experimental datasets. These methods are either based on Bayesian scores [Cooper and Herskovits, 1992, Heckerman et al., 1995] or rely on the identification of structural independencies, which correspond to missing edges in the underlying network [Pearl and Verma, 1991, Spirtes et al., 1993, 2000]. Bayesian inference methods typically require heuristic search strategies, such as hill-climbing algorithms, to sample the super-exponential space of possible networks. By contrast, constraint-based methods are expected to run in polynomial time on sparse underlying graphs, provided that a correct list of conditional independencies is available. Yet, in practice, conditional independencies need to be ascertained from the available observational data, based on adjustable statistical significance levels, and are not robust to sampling noise from finite datasets.

1.2.1 The state-of-the-art approaches

1.2.1.1 Constraint-based methods

Structure learning algorithms based on the identification of structural constraints, so-called constraint-based approaches such as the PC [Spirtes and Glymour, 1991] and IC [Pearl and Verma, 1991] algorithms, do not score and compare alternative networks. Instead they aim at ascertaining conditional independencies between variables to directly infer the Markov equivalent class of all causal graphs compatible with the available observational data. Yet, these methods are not robust to sampling noise in finite datasets as early errors in removing edges from the complete graph typically trigger the accumulation of compensatory errors later on in the pruning process. This cascading effect makes the constraint-based approaches sensitive to the adjustable significance level α, required for the conditional independence tests. In addition, traditional constraint-based methods are not robust to the order in which the conditional independence tests are processed, which prompted recent algorithmic improvements intending to achieve order-independence [Colombo and Maathuis, 2014] (see Chapter 3).

1.2.1.2 Search-and-score methods

By contrast, search-and-score methods [Chickering, 2002, Cooper and Herskovits, 1992, Friedman and Koller, 2003, Heckerman et al., 1995] have the advantage of allowing for quantitative comparisons between alternative networks through their Bayesian scores but they are limited to rather small causal graphs due to the super-exponential space of possible directed graphs to sample. Hence, search-and-score approaches typically require either suitable prior restrictions on the structures [Koivisto and Sood, 2004, Silander and Myllymaki, 2006] or heuristic strategies such as hill-climbing algorithms to sample network space [Bouckaert, 1994, Chickering et al., 1995, Friedman et al., 1999] (see Chapter 4).

1.2.2 3off2: a novel network reconstruction method

In the second and main part of my thesis, I have developed a more robust approach to reconstruct graphical models from finite datasets that combines the main advantages of constraint-based and search-and-score methods to reliably reconstruct graphical models despite inherent sampling noise in finite observational datasets.

This new hybrid method, embedded in an information theoretic framework (Chapter 5), confidently ascertain structural independencies in causal graphs based on the ranking of their most likely contributing nodes. In a nutshell, this local optimization scheme and corresponding 3off2 algorithm iteratively *take off* the most likely conditional 3-point information from the 2-point (mutual) information between each pair of nodes. Conditional independencies are thus derived by progressively collecting the most significant indirect contributions to all pairwise mutual information (Chapter 6, section 6.3.1). The resulting network skeleton is then partially directed by orienting and propagating edge directions, based on the sign and magnitude of the conditional 3-point information of unshielded triples (Chapter 6, section 6.3.2.2). The approach is shown to outperform both constraint-based and Bayesian inference methods on a range of benchmark networks (Chapter 7) and to retrieve more experimentally ascertained results than with other available methods in the field of systems biology (Chapters 8-11).

Cell Reports
Article

On the Expansion of "Dangerous" Gene Repertoires by Whole-Genome Duplications in Early Vertebrates

Param Priya Singh,[1,3] Séverine Affeldt,[1,3] Ilaria Cascone,[2] Rasim Selimoglu,[2] Jacques Camonis,[2] and Hervé Isambert[1,*]
[1]CNRS UMR168
[2]INSERM U830
UPMC, Institut Curie, Research Center, 26, rue d'Ulm, 75248 Paris, France
[3]These authors contributed equally to this work
*Correspondence: herve.isambert@curie.fr
http://dx.doi.org/10.1016/j.celrep.2012.09.034

SUMMARY

The emergence and evolutionary expansion of gene families implicated in cancers and other severe genetic diseases is an evolutionary oddity from a natural selection perspective. Here, we show that gene families prone to deleterious mutations in the human genome have been preferentially expanded by the retention of "ohnolog" genes from two rounds of whole-genome duplication (WGD) dating back from the onset of jawed vertebrates. We further demonstrate that the retention of many ohnologs suspected to be dosage balanced is in fact indirectly mediated by their susceptibility to deleterious mutations. This enhanced retention of "dangerous" ohnologs, defined as prone to autosomal-dominant deleterious mutations, is shown to be a consequence of WGD-induced speciation and the ensuing purifying selection in post-WGD species. These findings highlight the importance of WGD-induced nonadaptive selection for the emergence of vertebrate complexity, while rationalizing, from an evolutionary perspective, the expansion of gene families frequently implicated in genetic disorders and cancers.

INTRODUCTION

Just as some genes happen to be more "essential," owing to their deleterious loss-of-function or null mutations, some genes can be classified as more "dangerous," due to their propensity to acquire deleterious gain-of-function mutations. This is, in particular, the case for oncogenes and genes with autoinhibitory protein folds, whose mutations typically lead to constitutively active mutants with dominant deleterious phenotypes (Pufall and Graves, 2002).

Dominant deleterious mutations, that are lethal or drastically reduce fitness over the lifespan of organisms, must have also impacted their long term evolution on timescales relevant for genome evolution (e.g., >10–100 million years [MY]). In fact, dominant disease genes in humans have been shown to be under strong purifying selection (Furney et al., 2006; Blekhman et al., 2008; Cai et al., 2009). Yet, "dangerous" gene families implicated in cancer and severe genetic diseases have also been greatly expanded by duplication in the course of vertebrate evolution. For example, the single orthologous locus, *Ras85D* in flies and *Let-60* in nematodes, has been duplicated into three *RAS* loci in typical vertebrates, *KRAS*, *HRAS*, and *NRAS*, that present permanently activating mutations in 20%–25% of all human tumors, even though *HRAS* and *NRAS* have also been shown to be dispensable for mouse growth and development (Ise et al., 2000; Esteban et al., 2001).

While the maintenance of essential genes is ensured by their lethal null mutations, the expansion of dangerous gene families remains an evolutionary puzzle from a natural selection perspective. Indeed, considering that many vertebrate disease genes are phylogenetically ancient (Domazet-Loso and Tautz, 2008; Cai et al., 2009; Dickerson and Robertson, 2012), and that their orthologs also cause severe genetic disorders in extant invertebrates (Berry et al., 1997; Ciocan et al., 2006; Robert, 2010), it is surprising that dangerous gene families have been duplicated more than other vertebrate genes without known dominant deleterious mutations. While gene duplicates can confer mutational robustness against loss-of-function mutations, multiple copies of genes prone to gain-of-function mutations are expected to lead to an overall aggravation of a species' susceptibility to genetic diseases and thus be opposed by purifying selection.

Two alternative hypotheses can be put forward to account for the surprising expansion of dangerous gene families. Either, the propensity of certain genes to acquire dominant deleterious mutations could be a mere by-product of their presumed advantageous functions. In that case, only the overall benefit of gene family expansion should matter, irrespective of the mechanism of gene duplication. Alternatively, gene susceptibility to dominant deleterious mutations could have played a driving role in the striking expansion of dangerous gene families. But what could have been the selection mechanism?

In this article, we report converging evidences supporting the latter hypothesis and propose a simple evolutionary model to explain the expansion of such dangerous gene families. It is based on the observation that the majority of human genes prone to dominant deleterious mutations, such as oncogenes and genes with autoinhibitory protein folds, have not been duplicated through small scale duplication (SSD). Instead, the expansion of these dangerous gene families can be traced back to two rounds of whole-genome duplication (WGD), that occurred at the

Cell PRESS
Open ACCESS

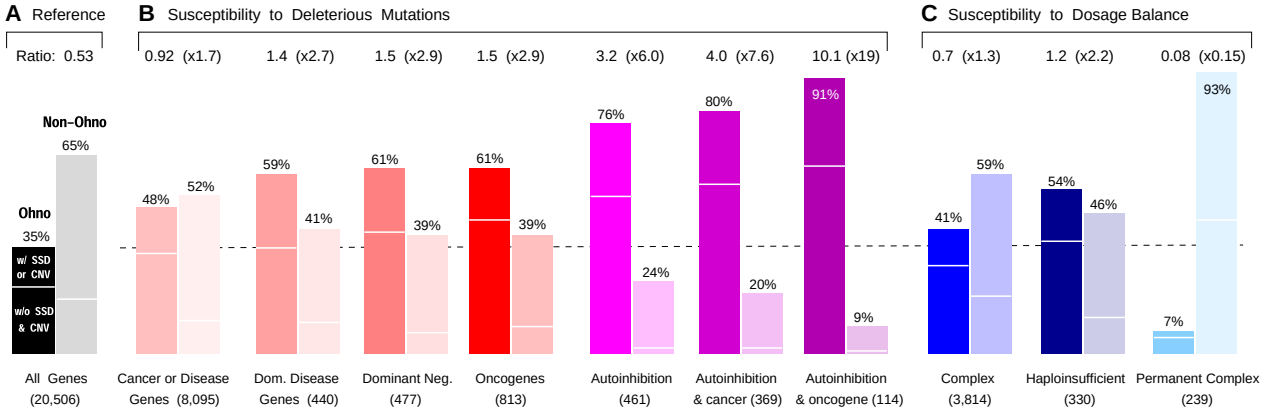


Figure 1. Prevalence of Retained Ohnologs in the Human Genome within Different Gene Classes

(A and B) Prevalence of retained ohnologs either "w/ SSD or CNV" or "w/o SSD & CNV" for all 20,506 human protein-coding genes (A), and gene classes susceptible to deleterious mutations (B). Note that gene classes with higher susceptibility to deleterious mutations retained more ohnologs.

(C) Ohnolog retention in gene classes susceptible to dosage balance constraints. Fold changes in ohnolog/nonohnolog ratios are given relative to the reference from all human genes in (A).

See also Figure S1.

onset of jawed vertebrates, some 500 MY ago (Ohno, 1970; Putnam et al., 2008).

These two rounds of WGD in the early vertebrate lineage are frequently credited with creating the conditions for the evolution of vertebrate complexity. Indeed, WGD-duplicated genes, so-called "ohnologs" in honor of Susumu Ohno (Ohno, 1970; Wolfe, 2000), have been preferentially retained in specific gene classes associated with organismal complexity, such as signal transduction pathways, transcription networks, and developmental genes (Maere et al., 2005; Blomme et al., 2006; Freeling and Thomas, 2006; Sémon and Wolfe, 2007; Makino and McLysaght, 2010; Huminiecki and Heldin, 2010). By contrast, gene duplicates coming from SSD are strongly biased toward different functional categories, such as antigen processing, immune response, and metabolism (Huminiecki and Heldin, 2010). SSD paralogs and WGD ohnologs also differ in their gene expression and protein network properties (Hakes et al., 2007; Guan et al., 2007). Furthermore, recent genome-wide analysis have shown that ohnologs in the human genome have experienced fewer SSD than "nonohnolog" genes and tend to be refractory to copy number variation (CNV) caused by polymorphism of small segmental duplications in human populations (Makino and McLysaght, 2010). These antagonist retention patterns of WGD and SSD/CNV gene duplicates in the human genome have been suggested to result from dosage balance constraints (Makino and McLysaght, 2010) on the relative expressions of multiple protein partners (Veitia, 2002), as proposed earlier for other organisms like yeast (Papp et al., 2003) and the paramecium (Aury et al., 2006).

In this article, we investigate the evolutionary causes responsible for the expansion of gene families prone to deleterious mutations in vertebrates and propose a simple evolutionary model accounting for their antagonistic retention pattern after WGD and SSD events. The retention of ohnologs in the human genome is shown to be more strongly associated with their susceptibility to deleterious mutations, than their functional importance or "essentiality." We also demonstrate using a causal inference analysis, that the retention of many ohnologs suspected to be dosage balanced is in fact an indirect effect of their higher susceptibility to deleterious mutations. We argue that the enhanced retention of dangerous ohnologs is a somewhat counterintuitive yet simple consequence of the speciation event triggered by WGD and the ensuing purifying selection in post-WGD species.

These findings rationalize, from an evolutionary perspective, the WGD expansion of gene families frequently implicated in genetic disorders, such as cancer, and highlight the importance of nonadaptive selection on the emergence of vertebrate complexity.

RESULTS

Genes Prone to Deleterious Mutations Retain More Ohnologs

We first analyzed a possible association between the susceptibility of human genes to deleterious mutations and their retention of ohnologs, as proposed in Gibson and Spring (1998) for multidomain proteins. To this end, we considered multiple classes of genes susceptible to deleterious mutations from experimentally verified databases and literature. These classes include cancer genes (from multiple sources including COSMIC [Forbes et al., 2011] and CancerGenes [Higgins et al., 2007]), genes mutated in other genetic disorders, dominant negative genes from OMIM, and genes with autoinhibitory protein folds (Experimental Procedures). We looked at the relative contributions of WGD and SSD in the expansion of these "dangerous" gene classes.

The results, depicted in Figures 1 and S1, demonstrate indeed a strong association between the retention of human ohnologs from vertebrate WGD and their reported susceptibility to deleterious mutations, as compared to nonohnologs, whereas an

opposite pattern is found for SSD/CNV gene duplicates. Overall, the 8,095 human genes associated with the occurrence of cancer and other genetic diseases have retained significantly more ohnologs than expected by chance, 48% versus 35% (48%; 3,844/8,095; $p = 1.3 \times 10^{-129}$, χ^2 test). Furthermore, these associations, which do not take into account the actual severity of the gene mutations, are clearly enhanced when the analysis is restricted to genes with direct experimental evidence of dominant deleterious mutations, such as dominant disease genes (59%; 261/440; $p = 1.7 \times 10^{-27}$, χ^2 test), dominant negative mutants (61%; 292/477; $p = 3.9 \times 10^{-34}$, χ^2 test), oncogenes (61%; 493/813; $p = 1.4 \times 10^{-54}$, χ^2 test), or genes exhibiting autoinhibitory constraints (76%; 350/461; $p = 2.7 \times 10^{-77}$, χ^2 test). The biased retention of ohnologs is even stronger for genes combining several factors associated with an enhanced susceptibility to deleterious mutations, such as cancer genes with autoinhibitory folds, (80%; 294/369; $p = 1.0 \times 10^{-73}$, χ^2 test), or oncogenes with autoinhibitory folds, (91%; 104/114; $p = 6.9 \times 10^{-37}$, χ^2 test).

This retention of dangerous ohnologs is illustrated on Table 1 that presents an up-to-date list of 76 hand-curated gene families of up to four ohnologs, exhibiting both autoinhibitory folds and oncogenic properties (see Table S1 for oncogenic and autoinhibitory details and references). These dangerous ohnologs are typically found along signal transduction cascades, from receptor tyrosine kinases and cytoplasmic or nuclear kinases to guanine exchange factors (GEF), GTPase activating proteins (GAP), and transcription factors (Table 1, gene classes A–E). In addition, autoinhibited oncogenes are also found in other ohnolog families with diverse functions (Table 1, gene class F). By contrast, we obtained a hand-curated list of only ten nonohnolog genes exhibiting both autoinhibitory and oncogenic properties, Table 1, gene class G (see Table S2 for oncogenic and autoinhibitory details and references). Interestingly, half of them (4/10) can be traced back to SSD events, which occurred after or at the same period of the two WGD in early vertebrate lineages (Table S2). All in all, this implies that >90% of known oncogenes with autoinhibitory folds have retained at least one ohnolog pair in the human genome (as well as, possibly, a few additional duplicates from more recent SSD events).

Ohnologs Are Conserved but More "Dangerous" than "Essential"

We then investigated whether the susceptibility of ohnologs to deleterious mutations could be directly quantified through comparative sequence analysis. We used Ka/Ks ratio estimates, which measure the proportion of nonsynonymous substitutions (Ka) to the proportion of synonymous substitutions (Ks) (Extended Results and Table S3). Ohnologs exhibit statistically lower Ka/Ks ratios, Figures 2, S2, and S3, which provides direct evidence of strong conservation, consistent with a higher susceptibility of ohnologs to deleterious mutations. Similar trends have also been reported for ohnologs specific to teleost fishes (Brunet et al., 2006) or to the more recent WGD in *Xenopus laevis* lineage (Sémon and Wolfe, 2008). Note, however, that the functional consequences of such deleterious mutations, leading either to a gain or a loss of function, cannot be directly inferred from Ka/Ks distributions. Yet, as outlined below, we found marked differences in the retention of "dangerous" ohnologs prone to dominant gain-of-function mutations and "essential" ohnologs exhibiting lethal loss-of-function or null mutations.

While autosomal-dominant disease genes exhibit a strong ohnolog retention bias (Figure 1B), 59% versus 35% (59%; 261/440; $p = 1.7 \times 10^{-27}$, χ^2 test), autosomal-recessive disease genes are not significantly enriched in ohnologs 37% versus 35% (37%; 221/598; $p = 0.24$, χ^2 test). Similarly, human orthologs of mouse genes, reported as being "essential" genes from large-scale null mutant studies in mouse, are not strongly enriched in ohnologs 56% versus 54% (56%; 1,537/2,729; $p = 3.8 \times 10^{-3}$, χ^2 test), where 54% = 3,190/5,956 is the global proportion of ohnologs among the 5,956 genes tested for null mutation in mouse (Experimental Procedures). In fact, this small enrichment becomes even nonsignificant once genes with dominant allelic mutants are removed from the list of 5,956 genes tested for essentiality in mouse, i.e., 50% versus 48% (50%; 760/1,525; $p = 0.09$, χ^2 test), where 48% = 1,782/3,739 is the global proportion of ohnologs among the 3,739 genes tested for essentiality in mouse, after removing dominant disease genes, oncogenes, and genes with dominant negative mutations or autoinhibitory folds.

All in all, this shows that the retention of ohnologs has been most enhanced for genes prone to autosomal-dominant deleterious mutations and not autosomal-recessive deleterious mutations. This suggests that the retention of ohnologs is more strongly related to their "dangerousness," as defined by their high susceptibility to dominant deleterious mutations, than their functional importance or "essentiality," as identified through large-scale null mutation studies in mouse.

Ultimately, we will argue that the "dangerousness" of ohnologs effectively controls their individual retention in the genomes of post-WGD species, as will be shown below in the section Model for the Retention of Dangerous Ohnologs.

Mixed Susceptibility of Human Ohnologs to Dosage Balance

An alternative hypothesis, focusing instead on the collective retention of interacting ohnologs, has been frequently invoked to account for the biased retention of ohnologs in unicellular organisms like yeast (Papp et al., 2003) or the paramecium (Aury et al., 2006) and in higher eukaryotes (Birchler et al., 2001; Makino and McLysaght, 2010).

This "dosage balance" hypothesis posits that interacting protein partners tend to maintain balanced expression levels in the course of evolution, in particular, for protein subunits of conserved complexes (Birchler et al., 2001; Veitia, 2002; Papp et al., 2003; Veitia, 2010; Makino and McLysaght, 2010). Thus, SSD of dosage balanced genes are thought to be generally detrimental through the dosage imbalance they induce, thereby raising the odds for their rapid nonfunctionalization (Papp et al., 2003; Maere et al., 2005). By contrast, rapid nonfunctionalization of ohnologs after WGD has been suggested to be opposed by dosage effect, in particular, for highly expressed genes and genes involved in protein complexes or metabolic pathways (Aury et al., 2006; Evlampiev and Isambert, 2007; Gout et al., 2010; Makino and McLysaght, 2010). This is because WGD initially preserves correct relative dosage between

Table 1. Ohnolog Families with Both Autoinhibitory and Oncogenic Properties

A. Ohnolog Receptor Tyrosine Kinases and Other Receptor Kinases									
ALK	*LTK*					*KIT*	*CSF1R*	*FLT3*	
EGFR	*ERBB2*	*ERBB3*	*ERBB4*			*MET*	*MST1R*		
FGFR1	*FGFR2*	*FGFR3*	*FGFR4*			*NPRA*	*NPRB*		
IGF1R	*INSR*	*INSRR*				*PDGFRA*	*PDGFRB*		
B. Ohnolog Cytoplasmic and Nuclear Protein Kinases									
ABL1	*ABL2*					*PKN1*	*PKN2*	*PKN3*	
ARAF	*BRAF*	*RAF1*				*PRKAA1*	*PRKAA2*		
AKT1	*AKT2*	*AKT3*				*PRKCA*	*PRKCB*	*PRKCG*	
CAMK1	*CAMK1D*	*CAMK1G*	*PNCK*			*PRKCE*	*PRKCH*		
CAMKK1	*CAMKK2*					*PRKCI*	*PRKCZ*		
CSNK1D	*CSNK1E*					*PRKD1*	*PRKD2*	*PRKD3*	
GSK3A	*GSK3B*					*PRKG1*	*PRKG2*		
GRK4	*GRK5*	*GRK6*				*PTK2*	*PTK2B*		
JAK1	*JAK2*	*JAK3*	*TYK2*			*RSK1*	*RSK2*	*RSK3*	*RSK4*
SRC	*FGR*	*FYN*	*YES1*			*MSK1*	*MSK2*		
HCK	*LCK*	*BLK*	*LYN*			*NDR1*	*NDR2*		
MKNK1	*MKNK2*					*SYK*	*ZAP70*		
NEK6	*NEK7*								
C. Ohnolog GEF									
ARHGEF3	*NET1*					*RALGDS*	*RGL1*	*RGL2*	*RGL3*
ARHGEF6	*COOL1*					*SOS1*	*SOS2*		
DBL	*DBS*	*MCF2L2*				*TIAM1*	*TIAM2*		
FGD1	*FGD2*	*FGD3*	*FGD4*			*TIM*	*WGEF*	*SGEF*	*NGEF*
PDZ-RHOGEF	*LSC*	*LARG*				*VAV1*	*VAV2*	*VAV3*	
P114-RHOGEF	*GEF-H1*								
D. Ohnolog GAP									
ASAP1	*ASAP2*	*ASAP3*				*PLXNA1*	*PLXNA2*	*PLXNA3*	*PLXNA4*
IQGAP1	*IQGAP2*	*IQGAP3*				*PLXNB1*	*PLXNB2*	*PLXNB3*	*PLXND1*
E. Ohnolog DNA Binding and Transcription Factors									
CEBPA	*CEBPB*	*CEBPE*				*IRF4*	*IRF8*	*IRF9*	
CUX1	*CUX2*					*MEIS1*	*MEIS2*	*MEIS3*	
ELK1	*ELK3*	*ELK4*				*p53*	*p63*	*p73*	
ETS1	*ETS2*					*RUNX1*	*RUNX2*	*RUNX3*	
ETV1	*ETV4*	*ETV5*				*SOX1*	*SOX2*	*SOX3*	
ETV6	*ETV7*								
F. Other Ohnolog Genes with Both Autoinhibitory and Oncogenic Properties									
ANP32A	*ANP32B*	*ANP32E*				*nNOS*	*eNOS*		
ATP2B1	*ATP2B2*	*ATP2B3*	*ATP2B4*			*NOTCH1*	*NOTCH2*	*NOTCH3*	
cIAP1|2	*XIAP*					*PLCB1*	*PLCB2*	*PLCB3*	
CCNT1	*CCNT2*					*PLCD1*	*PLCD3*	*PLCD4*	
FLNA	*FLNB*	*FLNC*				*PLCG1*	*PLCG2*		
FURIN	*PCSK4*					*PTPN1*	*PTPN2*		
KPNA2	*KPNA7*					*SMURF1*	*SMURF2*		
NEDD4	*NEDD4L*					*TRPV1|3*	*TRPV2*	*TRPV4*	*TRPV5|6*
NOXA1	*NOXA2*								
G. Nonohnolog Genes with Both Autoinhibitory and Oncogenic Properties									
CAMK4	*ELF3*	*MELK*	*MOS*	*PDPK1*	*BRK*	*PTPN11*	*RET*	*RPS6KB1*	*TTN*

GEF, guanine exchange factors; GAP, GTPase activating proteins.
See also Tables S1 and S2.

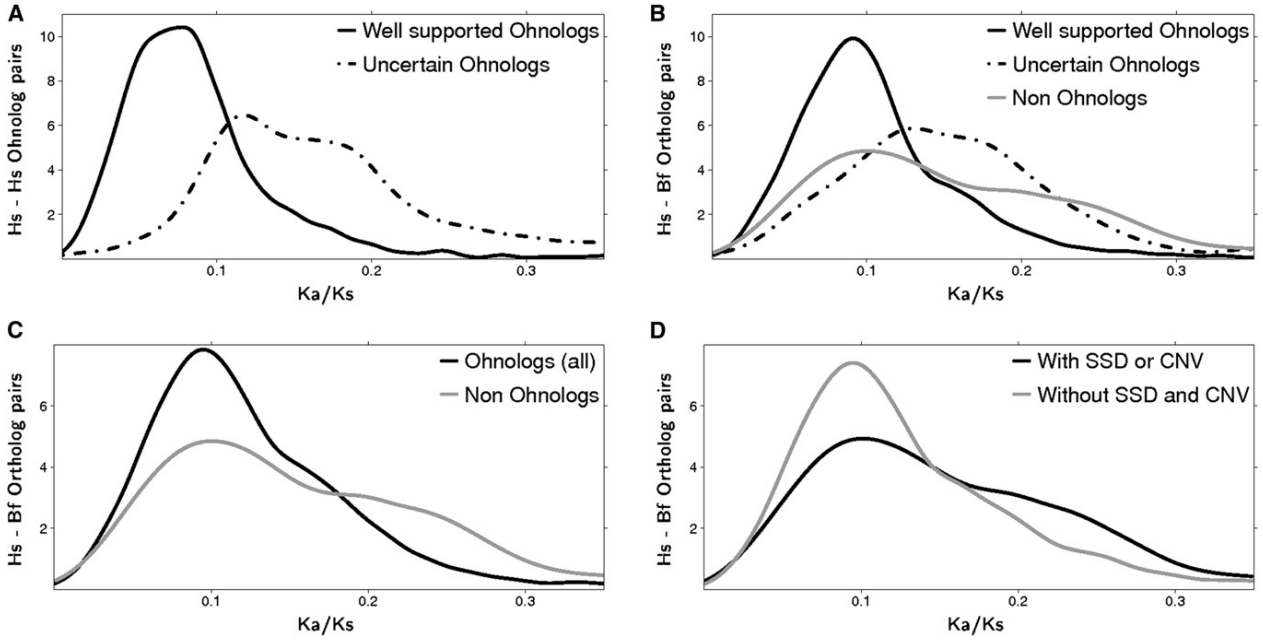


Figure 2. Ka/Ks Distributions for WGD and SSD or CNV Duplicates in the Human Genome

(A–D) Ka/Ks distributions for human-human (Hs-Hs) ohnolog pairs (A) and human-amphioxus (Hs-Bf) ortholog pairs (B) with different confidence status (see Extended Results). Ka/Ks distributions for human-amphioxus (Hs-Bf) ortholog pairs involving a human ohnolog (C) and for human-amphioxus (Hs-Bf) ortholog pairs exhibiting either SSD or CNV (D).

See also the Extended Results, Figures S2 and S3, and Table S3 for statistical significance and comparison with other invertebrate outgroups.

expressed genes, whereas subsequent random nonfunctionalization of individual ohnologs disrupts this initial dosage balance. For instance, yeast *Saccharomyces cerevisiae* has retained 76% of its ribosomal gene ohnologs from a 150 MY old WGD (Kellis et al., 2004; Lin et al., 2007), although the maintenance of these ohnologs has been suggested to require frequent gene conversion events (Kellis et al., 2004; Evangelisti and Conant, 2010) as well as fine-tuned dosage compensation to ensure a balanced expression with the remaining 24% ribosomal genes having lost their ohnologs (Zeevi et al., 2011).

Following on this dosage balance hypothesis, we performed statistical analysis on multiprotein complexes from HPRD (Keshava Prasad et al., 2009) and CORUM (Ruepp et al., 2010) databases and a hand-curated list of permanent complexes (Zanivan et al., 2007) (Experimental Procedures) to investigate for a possible association between the retention of human ohnologs and their susceptibility to dosage balance constraints.

The results depicted in Figure 1C demonstrate, in agreement with (Makino and McLysaght, 2010), that genes implicated in multiprotein complexes have retained significantly more ohnologs than expected by chance, 41% versus 35% (41%; 1,567/3,814; $p = 8.7 \times 10^{-17}$, χ^2 test). This trend is also enhanced when focusing on haploinsufficient genes, that are known for their actual sensitivity to dosage balance constraints (Qian and Zhang, 2008) (54%; 179/330; $p = 8.0 \times 10^{-14}$, χ^2 test).

Yet, surprisingly, an opposite trend corresponding to the elimination of ohnologs is observed for genes implicated in permanent complexes, that are presumably strongly sensitive to dosage balance constraints (7.5%; 18/239; $p = 1.2 \times 10^{-18}$, χ^2 test) (Figure 1C). In fact, looking more closely at the few human ohnologs, that have not been eliminated from permanent complexes (Table 2), we found that they are likely under less stringent dosage balance constraints than most proteins in permanent complexes, as they typically coassociate with mitochondrial proteins or form large multimeric subcomplexes with intrinsic stoichiometry disequilibrium.

This suggests that the elimination of most ohnologs from permanent complexes is, in fact, strongly favored under dosage imbalance and becomes likely inevitable once a few of those ohnologs have been accidentally lost following WGD. Indeed, the uneven elimination of ohnologs in permanent complexes is expected to lead to the assembly of nonfunctional, partially formed complexes detrimental to the cell, unless dosage compensation mechanisms effectively re-establish proper dosage balance at the level of gene regulation (Birchler et al., 2001), as for yeast ribosomal proteins (Zeevi et al., 2011). By contrast, transient complexes, which are typically more modular than permanent complexes, are expected to accommodate such dosage changes more easily, as they do not usually require the same strict balance in the expression levels of their protein partners.

These findings on the differences in retention of human ohnologs between permanent and more transient complexes suggest the relevance of different underlying causes. Although dosage balance presumably remains the primary evolutionary constraint in permanent complexes (<2% of human genes), which lead to the elimination of ohnologs in permanent complexes in

Table 2. Low Retention of Ohnologs in Permanent Complexes

Permanent Complexes[a]	Number of Ohnologs	Intrinsic Stoichiometry Disequilibrium of Ohnologs in Permanent Complexes
ATP F0	3/12	the 3 ohnologs ATP5G1-3 form the 10-mer C-ring of the F-type ATP synthase
ATP F1	0/5	
COX	2/11	the 2 ohnologs COX4I1,2 coassemble with 3 mitochondrial encoded genes
SRS	2/32	Ohnologs are X-linked RPS4X (with no X-inactivation) and Y-linked RPS4Y1
Mitochondrial SRS	0/30	
LRS	2/50	RPL3 and RPL39 have ohnologs RPL3L and RPL39L with unknown functions
Mitochondrial LRS	0/48	
Proteasome	2/31	ohnologs PSMA7 or PSMA7L are included in the 2 rings of 7 α subunits
Pyruvate dehydrogenase	0/5	
RNA Pol II	0/12	
RNA Pol III	0/9	

COX, cytochrome c oxidase; LRS, large ribosomal subunit; SRS, small ribosomal subunit.
[a]Zanivan et al., 2007.

vertebrate genomes, gene susceptibility to deleterious mutations may be more relevant for the retention of ohnologs within the 17% of human genes participating in more transient complexes. For instance, transient complexes involved in phosphorylation cascades or GTPase signaling pathways are known to be more sensitive to the level of activation of their protein partners than to their total expression levels. Thus, although the active forms of multistate proteins typically amount to a small fraction of their total expression level, hence providing a large dynamic range for signal transduction, it also makes them particularly susceptible to gain-of-function mutations. Such mutations can shift protein activation levels 10- to 100-fold without changes in expression levels and likely underlie stronger evolutionary constraints than the 2-fold dosage imbalance caused by gene duplication.

Indirect Cause of Ohnolog Retention in Protein Complex

To further investigate the relative effects of dosage balance and gene susceptibility to deleterious mutations, we analyzed whether the overall enhanced retention of ohnologs within multiprotein complexes (Figure 1C) could indirectly result from an enhanced susceptibility to deleterious mutations. Indeed, as outlined in Figure 3A, cancer and disease genes are more prevalent within complexes than expected by chance, 29% versus 19% (29%; 2,362/8,095; $p = 3.7 \times 10^{-132}$, χ^2 test) and this trend is enhanced for genes with stronger susceptibility to deleterious mutations, such as oncogenes (39%; 320/813; $p = 2.9 \times 10^{-52}$, χ^2 test) or oncogenes with autoinhibitory folds (59%; 67/114; $p = 2.9 \times 10^{-28}$, χ^2 test). By contrast, ohnologs are only slightly, although significantly, more prevalent in complexes than expected by chance, 22% versus 19% (22%; 1,567/7,110; $p = 9.0 \times 10^{-14}$, χ^2 test), whereas the proportion implicated in cancer or disease genes is clearly enhanced 54% versus 39% (54%; 3,844/7,110; $p = 9.5 \times 10^{-140}$, χ^2 test).

To go beyond these simple statistical associations and quantify the direct versus indirect effects of deleterious mutations and dosage balance constraints on the biased retention of human ohnologs, we have performed a Mediation analysis following the approach of Pearl (Pearl, 2001, 2011). The Mediation framework, developed in the context of causal inference analysis, aims at uncovering, beyond statistical correlations, causal pathways along which changes in multivariate properties are transmitted from a cause, *X*, to an effect, *Y*. More specifically, a Mediation analysis assesses the importance of a mediator, *M*, in transmitting the indirect effect of *X* on the response $Y \equiv Y(x,m(x))$ (Figure 3B).

Mediation analyses have been typically used in social sciences research (Baron and Kenny, 1986) as, for instance, in the context of legal disputes over alleged discriminatory hiring. In such cases, the problem is to establish that gender or race (*X*) have directly influenced hiring (*Y*) and not simply indirectly through differences in qualification or experience (*M*). Mediation analyses have also been used in epidemiology, as in a formal study (Robins and Greenland, 1992) that establishes the direct effect of smoking (*X*) on the incidence of cardiovascular diseases (*Y*), while taking into account the indirect effect of other aggravating factors, such as hyperlipidemia (*M*).

In this report, we have applied the Mediation analysis to genomic data to discriminate between direct effect (*DE*) and indirect effect (*IE*) of deleterious mutations (*X* or *M*) and dosage balance constraints (*M* or *X*) on the biased retention of human ohnologs (*Y*). The results, derived in Extended Experimental Procedures (Table S4) and summarized in Figure 3C and Table S5, demonstrate that the retention of ohnologs in the human genome is more directly caused by their susceptibility to deleterious mutations than their interactions within multiprotein complexes.

Indeed, the direct causal effect of a change from "noncomplex" to "complex" proteins only accounts for 23% of a small total effect (*TE*) of complex on the retention of ohnologs (*DE*/*TE* = 23% with *TE* = 0.079), whereas 82% of this small total effect is indirectly mediated by their susceptibility to deleterious mutations (*IE*/*TE* = 82% with 5% nonlinear coupling between direct and indirect effects) (Extended Results). By contrast, the alternative hypothesis, assuming a direct effect of deleterious mutations, accounts for 99% of a three times larger total effect on ohnolog retention (*DE*/*TE* = 99% with *TE* = 0.23), whereas the "complex" versus "noncomplex" status of human genes

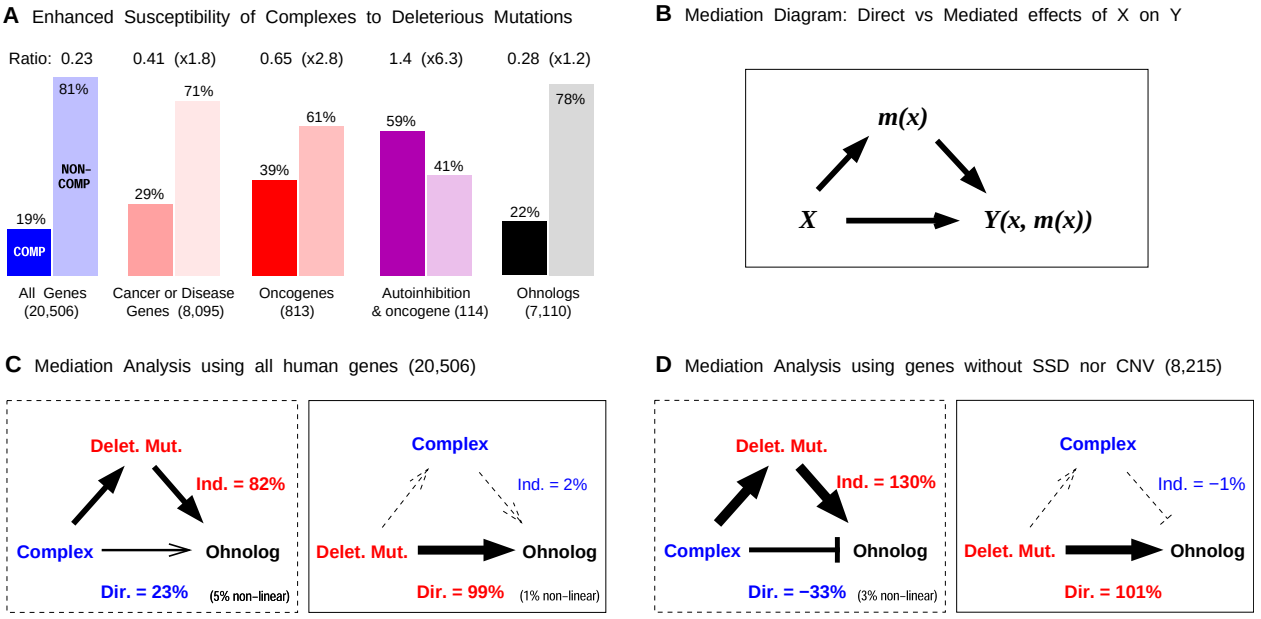


Figure 3. Mediation Analysis of the Indirect Effect of Deleterious Mutations on the Retention of Ohnologs in Multiprotein Complexes

(A) Enhanced susceptibility of complexes to deleterious mutations.

(B) Mediation diagram depicting the direct versus indirect (i.e., mediated) effects of the cause *X* on the outcome *Y*(*x*,*m*(*x*)) (Pearl, 2011). See also Extended Experimental Procedures.

(C and D) Quantitative Mediation analysis of direct versus indirect effects of deleterious mutations and dosage balance on the retention of human ohnologs using (C) all human genes (20,506) or (D) all human genes without SSD nor CNV (8,215). The thickness of the arrows outlines the relative importance of the corresponding direct or indirect effects. These results are consistent with those obtained from partial correlation analysis.

See also the main text, Extended Results, and Tables S4, S5, and S6.

has a negligible indirect effect on ohnolog retention in this case (*IE*/*TE* = 2%) (Extended Results). These trends are also further enhanced when the analysis is restricted to the 40% of human genes (8,215) without SSD and CNV duplicates (Figure 3D; Table S5; Extended Results). In fact, the direct effect of multiprotein complexes then tends to oppose the retention of ohnologs (*DE*/*TE* = −33% with *TE* = 0.064), as in the case of permanent complexes detailed above, but on an increased sample size of 8,215 genes without SSD or CNV duplicates (i.e., more than a third of human genes) in place of 239 genes from permanent complexes. By contrast, there is a five times larger total effect due to the direct effect of deleterious mutations on the retention of ohnologs (*DE*/*TE* = 101% with *TE* = 0.32), Figure 3D. This is an instance of Simpson's paradox, where two effects oppose each other, thereby, revealing the existence of conflicting underlying causes, namely, a strong positive effect of deleterious mutations and a small negative effect of dosage balance constraints on the retention of human ohnologs without SSD and CNV duplicates.

We have also examined the effects of other alternative properties on the retention of ohnologs (Extended Results; Table S5). In particular, we have found that gene expression levels and Ka/Ks ratios do not significantly mediate the effect of deleterious mutations on the retention of ohnologs. In fact, gene expression levels (Extended Experimental Procedures) have a negligible total effect on the retention of human ohnologs (*TE* = 0.003), by contrast to what has been reported for the paramecium (Gout et al., 2009). The total effects of Ka/Ks on ohnolog retention are also lower than the total effects of deleterious mutations, as *TEs* from deleterious mutations are ~2- to 3-fold stronger than *TEs* from Ka/Ks and become >10-fold stronger for genes without SSD and CNV (Extended Results).

In addition, we have performed a complementary systematic study of all these genomics properties using partial correlation analysis, which aims at "removing" the effect of a third property (*Z*) on the standard pair correlations between two variables (*X*) and (*Y*). The results detailed in Extended Results and Table S6 are entirely consistent with those obtained through mediation analysis, although the two approaches are not equivalent. Indeed, although mediation effects require partial correlation, partial correlation does not imply mediation, in general (Extended Results).

All in all, these results support the fact that the retention of ohnologs in the human genome is more strongly associated with their "dangerousness" (i.e., susceptibility to dominant deleterious mutations) than with their functional importance ("essentiality"), sensitivity to dosage balance, absolute expression levels or sequence conservation (i.e., Ka/Ks).

Model for the Retention of "Dangerous" Ohnologs

As demonstrated above, human genes with a documented sensitivity to dominant deleterious mutations have retained statistically more ohnologs from the two WGD events at the

Cell PRESS
Open ACCESS

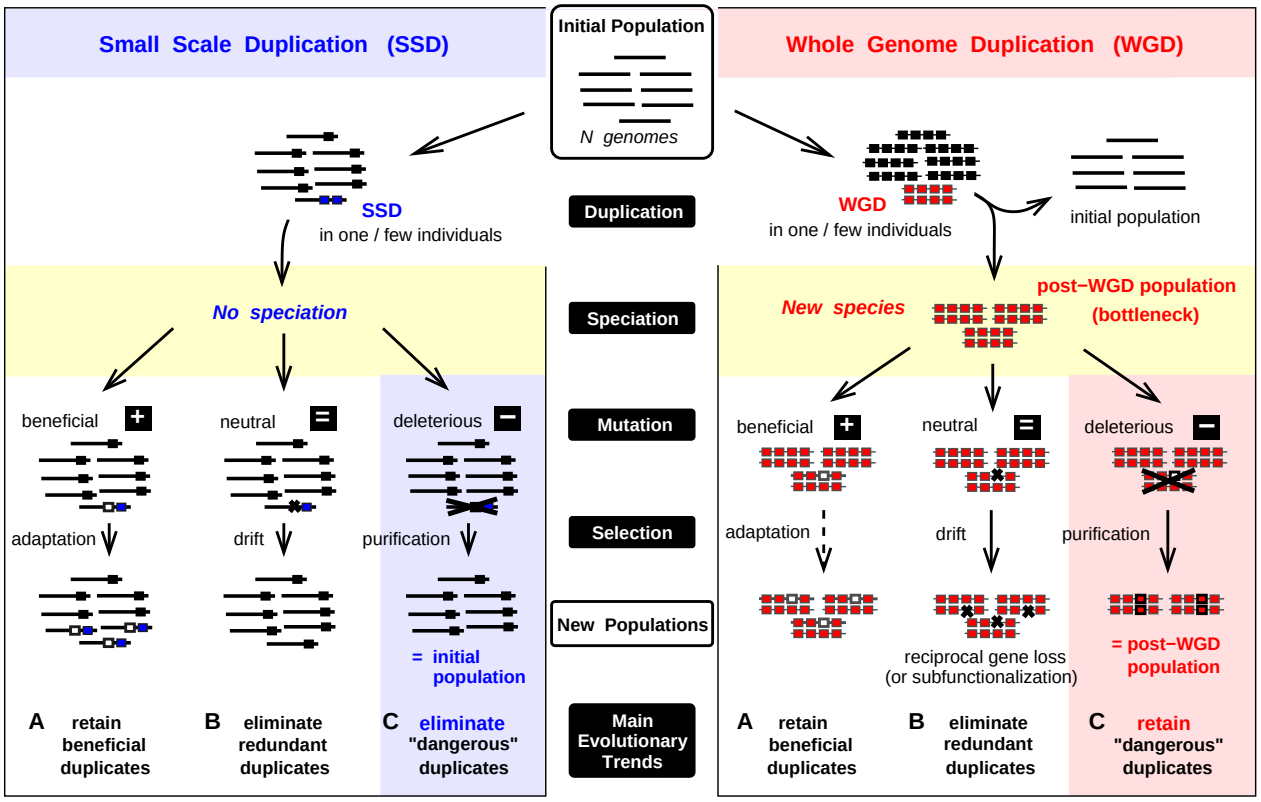


Figure 4. Evolutionary Trends of Duplicated Genes following SSD or WGD

(A–C) Horizontal lines represent the genome of different individuals. Square blocks symbolize the genes, duplicated (SSD: blue; WGD: red) or not (black). Black crosses highlight the loss of one gene (small crosses) or the elimination of an individual (larger crosses), whereas bordered square blocks emphasize retained mutated copies. Evolutionary scenarios are depicted at the population genetics level following either a SSD (left panel) or a WGD (right panel) in one or a few individuals of an initial population. Unlike SSD, WGD is invariably coupled to a speciation event, owing to the difference in ploidy between pre- and post-WGD individuals. Three possible scenarios—beneficial (A), neutral or nearly neutral (B), or deleterious mutations (C) in one gene duplicate—are outlined in post-SSD and post-WGD populations. The main difference concerns the mutation/selection process of "dangerous" genes, i.e. genes prone to autosomal-dominant deleterious mutations (C). See main text for a detailed description.

onset of jawed vertebrates. This suggests that ohnologs have been retained in vertebrate genomes, not because they initially brought selective advantages following WGD, but because their mutations were more likely detrimental or lethal than nonfunctional, thereby preventing their rapid elimination from the genomes of surviving individuals following WGD transitions, as outlined in the evolutionary model depicted in Figure 4.

For completeness and clarity, Figure 4 examines all possible evolutionary scenarios following either a SSD or a WGD duplication event in the genome of one or a few individuals in an initial population. The first and critical difference between SSD and WGD duplication events occurs at the population genetics level with an obligate speciation following WGD event, owing to the difference in ploidy between pre- and post-WGD individuals. As a result, all individuals in the post-WGD population carry twice as many genes as their pre-WGD relatives, whereas only a few individuals in the post-SSD population carry a single small duplicated region. Figure 4 then outlines the three mutation/selection scenarios focusing on a single gene duplicate in the genomes of post-SSD or post-WGD populations: (A) Beneficial mutations after SSD or WGD are expected to spread and become eventually fixed in the new populations, although the bottleneck in population size following WGD limits in practice the efficacy of adaptation in post-WGD species. (B) Neutral or nearly neutral mutations mainly lead to the random nonfunctionalization of one copy of most redundant gene duplicates and, therefore, to their elimination following both SSD and WGD events. In post-WGD populations, this results in the "reciprocal gene loss" of most gene duplicates, which is also known to lead to further speciations in post-WGD species, owing to the interbreeding incompatibility between post-WGD individuals with different "reciprocal gene loss" pattern (Lynch and Force, 2000a). Alternatively, neutral or nearly neutral mutations can also result in the eventual retention of both duplicate copies through subfunctionalization (Hughes, 1994; Lynch and Force, 2000b), that is, by rendering each duplicate copy unable to perform all the functions of their ancestral gene (see Discussion). (C) Finally, dominant deleterious mutations favor the elimination of the individuals

(or their descendants) harboring them through purifying selection. However, this typically leads to opposite outcomes in post-SSD and post-WGD populations. In post-SSD populations, dominant deleterious mutations will tend to eliminate SSD duplicates before they have the time to reach fixation (see below). By contrast, in post-WGD populations, where all ohnologs have been initially fixed through WGD-induced speciation, purifying selection will effectively favor the retention of dangerous ohnologs, as all surviving individuals still present (nondeleterious) functional copies of these dangerous genes.

Note, in particular, that this somewhat counterintuitive evolutionary model for the retention of "dangerous" ohnologs hinges on two unique features:

(1) It requires an autosomal dominance of deleterious mutations, in agreement with our observation, above, that retained ohnologs are more "dangerous" than "essential."
(2) It relies on the fact that successful WGD events start with a concomitant speciation event, which immediately fixes all ohnolog duplicates in the initial post-WGD population (Figure 4).

Note, also, that the same evolutionary trend is expected for dangerous SSD duplicates that would have the time (t) to become fixed through genetic drift in a population of size N before deleterious mutations can arise at a rate K, i.e., $t = 4N < 1/K$. This corresponds to a population bottleneck effect with $N < 1/(4K) \approx 5{,}000–10{,}000$ for typical vertebrates.

DISCUSSION

Beyond human and vertebrate genomes, WGD events have now been established in all major eukaryote kingdoms (Sémon and Wolfe, 2007; Evlampiev and Isambert, 2007). Unlike SSD events, WGD transitions provide a unique evolutionary mechanism, enabling the simultaneous duplication of entire genetic pathways and multiprotein complexes, followed by long periods of functional divergence and extensive loss of ohnologs (Aury et al., 2006). Moreover, although both WGD and SSD events have expanded the gene repertoires and resulting protein networks (Evlampiev and Isambert, 2007; Evlampiev and Isambert, 2008) of eukaryotes, it has become increasingly clear that WGD and SSD events actually lead to the expansion of different gene classes in the course of evolution, (Maere et al., 2005; Aury et al., 2006; Sémon and Wolfe, 2007; Makino and McLysaght, 2010; Huminiecki and Heldin, 2010; and this study).

In this article, we report that WGD have effectively favored the expansion of gene families prone to deleterious mutations in the human genome, such as genes implicated in cancer and genes with autoinhibitory interactions. In particular, we found that the retention of many ohnologs suspected to be dosage balanced is in fact indirectly mediated by their susceptibility to deleterious mutations.

From a broader perspective, a number of studies have now shown that many genomic properties, such as gene essentiality, duplicability, functional ontology, network connectivity, expression level, mutational robustness, divergence rates, etc., all appear to be correlated to some extent. In the light of the present study, we expect that many of these statistically significant correlations mainly result from indirect rather than direct associations, which may even frequently oppose each other. This highlights the need to rely on more advanced inference methods to analyze the multiple, direct, and indirect causes underlying the evolution of specific gene repertoires.

In the present study, we have quantitatively analyzed the direct versus indirect effects of the susceptibility of human genes to deleterious mutation and dosage balance constraints on the retention of ohnologs and proposed a simple evolutionary mechanism to account for the initial retention of "dangerous" ohnologs after WGD (Figure 4). On longer timescales, we expect that this initial retention bias of "dangerous" ohnologs effectively promote a prolonged genetic drift and, thus, a progressive functional divergence between ohnolog pairs. This eventually favors the subfunctionalization (Hughes, 1994; Lynch and Force, 2000b) of ancestral functions between ohnolog pairs, which ultimately warrants their long-term maintenance following WGD events.

Note, however, that this subfunctionalization process requires that the expression of ohnologs is not rapidly suppressed by large-scale deletion or silencing mutations in regulatory regions. As ohnolog pairs are not arranged in tandem, large-scale deletions through unequal crossing-over cannot typically remove entire ohnolog duplicates while preserving the integrity of nearby genes. Furthermore, as the size of promoter or enhancer regions is typically much smaller than UTRs and coding regions, one expects that the rate of transcriptional silencing does not exceed the rates of functional silencing and divergence in UTRs and coding regions. In fact, early estimates (Nadeau and Sankoff, 1997) showed that gene loss and functional divergence after genome duplications in early vertebrates occurred at comparable rates in gene families including at least two ohnologs. This is also directly evidenced by pseudotetraploid species like the vertebrate *Xenopus laevis*, which still retains ~40% of its initial ohnologs from a 30-million-year-old WGD (Sémon and Wolfe, 2008). All in all, this suggests that ohnologs prone to dominant deleterious mutations have at least a few million years to diverge and become nonredundant genes before they have a chance to be deleted or transcriptionally silenced.

Yet, we found that the retention of these dangerous ohnologs remains intrinsically stochastic by nature as many of them have also been eliminated following WGD events. This presumably occurred through loss-of-function mutations, transcriptional silencing, or large-scale deletion before ohnolog pairs could diverge and become nonredundant genes. More quantitatively, a simple theoretical estimate, derived from the long-term retention statistics of Figure 1, shows that only 6%–10% of the initial ohnolog duplicates have been retained on average at each round of WGD, Figure 5 (see Extended Results for details). By comparison, ~23%–30% of the initial ohnologs prone to gain-of-function mutations have been retained on average at each WGD. This implies that genes susceptible to deleterious mutations are two to five times more likely to retain ohnologs on long evolutionary timescales. Moreover, genes combining several factors associated with enhanced susceptibility to autosomal-dominant deleterious mutations are shown to be more than ten times more

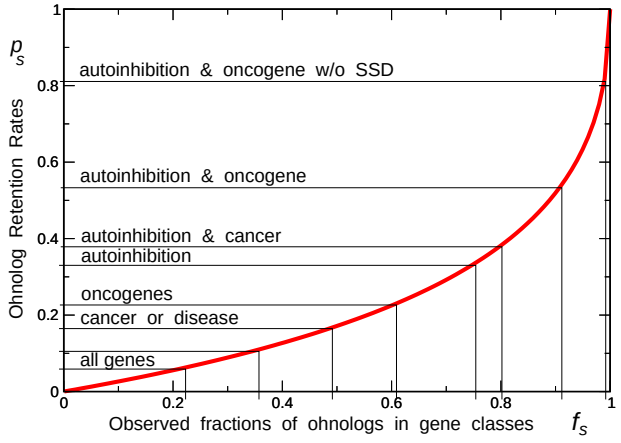


Figure 5. Estimates of Ohnolog Retention Rates

Estimates of ohnolog retention rates p_s in early vertebrates from the observed fraction f_s of ohnologs in the human genome for gene classes, s, with increasing susceptibility to deleterious mutations. The theoretical estimate (red curve) is obtained assuming that the retentions of ohnologs were comparable for each of the two WGD at the onset of vertebrates, and reads $P_s = 2/f_s - 1 - \sqrt{(2/f_s - 1)^2 - 1}$ as detailed in the Extended Results and Tables S7 and S8.

likely to retain ohnologs than genes lacking gain-of-function mutations (Figure 5), as illustrated on the examples of oncogenes with autoinhibitory folds (Table 1).

In turn, the elimination of ohnologs has been shown to drive further speciation events within post-WGD (sub)populations, due to the emergence of recombination barriers from the accumulation of differences in ohnolog deletion patterns between post-WGD individuals (Lynch and Force, 2000a). The resulting fragmentation of post-WGD subpopulations is then expected to sustain negative selection pressure that favors the retention of the remaining ohnolog pairs prone to deleterious mutations, as outlined in Figure 4. Hence, although most WGDs are unlikely to bring much fitness benefit on short evolutionary timescales (if only due to the population bottlenecks associated with WGD-induced speciations; Figure 4), they provide a unique evolutionary mechanism to experiment virtually unlimited combinations of regulation/deletion patterns from redundant ohnolog genes. Over long timescales (>100–500 MY), such trial and error combinations have visibly led to the evolutionary success and radiation of WGD species.

In summary, we present evidence supporting an evolutionary link between the susceptibility of human genes to dominant deleterious mutations and the documented expansion of these "dangerous" gene families by two WGD events at the onset of jawed vertebrates. We propose that deleterious mutations, responsible for many cancers and other severe genetic diseases on the lifespan of human individuals, have also underlain purifying selection over long evolutionary timescales, which effectively favored the retention of vertebrate ohnologs prone to dominant deleterious mutations, as outlined in Figure 4. From a population genetics perspective, we argue that this counterintuitive retention of dangerous ohnologs hinges in fact on WGD-induced speciation events, which are largely credited for the genetic complexity and successful radiation of vertebrate species.

These findings highlight the importance of purifying selection from WGD events on the evolution of vertebrates and, beyond, exemplify the role of nonadaptive forces on the emergence of eukaryote complexity (Fernández and Lynch, 2011).

EXPERIMENTAL PROCEDURES

WGD Duplicated Genes or "Ohnologs"

Human ohnolog genes were obtained from (Makino and McLysaght, 2010). Makino and McLysaght compared different vertebrate and six nonvertebrate outgroup genomes to identify ohnologs in the human genome. The final data set consists of 8,653 ohnolog pairs and 7,110 unique ohnologs. We further divided ohnologs into well supported (3,963), plausible (894), and more uncertain (2,253) ohnologs (see Extended Experimental Procedures).

SSD Duplicated Genes

We identified paralogous genes within the human genome from sequence similarity search. We obtained a total of 11,185 SSD genes. In particular, paralogs that were not annotated as ohnologs were taken to be SSD genes (see Extended Experimental Procedures).

Genes with CNV

CNV regions were obtained from Database of Genomic Variants (Zhang et al., 2006). A total of 5,709 genes were identified to be CNV genes as their entire coding sequence fell within one of the CNV regions.

Cancer and Disease Genes

We obtained cancer genes from multiple databases, including COSMIC (Forbes et al., 2011) and CancerGenes (Higgins et al., 2007), listed in Table S7. The detailed list of 6,917 cancer genes is given in Table S8 with a hand-curated list of 813 verified or predicted (Bozic et al., 2010) oncogenes (see Extended Experimental Procedures). We obtained 2,580 disease genes from the "Morbidmap" database of OMIM and hand curated subsets of 440 autosomal-dominant and 598 autosomal-recessive disease genes from Blekhman et al. (2008).

Genes with Autoinhibitory Folds

To obtain genes coding for proteins with autoinhibitory folds we searched PubMed with keyword "autoinhibitory domain" and retrieved relevant autoinhibitory genes and domains manually. Further gene candidates with autoinhibitory folds were obtained from databases, OMIM, SwissProt, NCBI Gene, and GeneCards using the parsing terms: auto/self-inhibit*. Careful manual curation of this list of gene candidates with the available literature finally yielded a total of 461 genes with autoinhibitory folds (94% of initial candidates).

Essential Genes

Mouse essential genes were obtained from Mouse Genome Informatics database. Essential genes were defined as genes having lethal or infertility phenotypes on loss-of-function or knockout mutations (2,729 genes) (see Extended Experimental Procedures).

Genes in Complexes and Permanent Complexes

Protein complexes were obtained from Human Protein Reference Database (HPRD) (Keshava Prasad et al., 2009) and CORUM database (Ruepp et al., 2010). In addition, a manually curated data set of permanent complexes (239 genes) was obtained from Zanivan et al. (2007). The final data set consists of 3,814 protein complex genes (see Extended Experimental Procedures).

Haploinsufficient and Dominant Negative Genes

Haploinsufficient and dominant negative candidate genes were obtained from parsing OMIM text files with Perl regular expressions. The resulting gene lists were manually curated with the available literature, yielding a total of

330 haploinsufficient genes (80% of initial candidates) and 477 dominant-negative genes (63% of initial candidates).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Results, Extended Experimental Procedures, three figures, and eight tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2012.09.034.

LICENSING INFORMATION

ACKNOWLEDGMENTS

P.P.S. acknowledges a PhD fellowship from Erasmus Mundus and Université Pierre et Marie Curie; S.A. acknowledges a PhD fellowship from Ministry of Higher Education and Research, France; I.C. acknowledges postdoctoral support from ANR (grant ANR-08-BLAN-0290); R.S. acknowledges PhD fellowships from INCa and ARC; H.I. and J.C. acknowledge funding from Foundation Pierre-Gilles de Gennes. We thank H. Roest Crollius, L. Peliti, S. Coscoy and V. Hakim for discussions.

Received: April 12, 2012
Revised: September 17, 2012
Accepted: September 27, 2012
Published: November 15, 2012

REFERENCES

Aury, J.M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., et al. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature *444*, 171–178.

Baron, R.M., and Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J. Pers. Soc. Psychol. *51*, 1173–1182.

Berry, L.W., Westlund, B., and Schedl, T. (1997). Germ-line tumor formation caused by activation of glp-1, a Caenorhabditis elegans member of the Notch family of receptors. Development *124*, 925–936.

Birchler, J.A., Bhadra, U., Bhadra, M.P., and Auger, D.L. (2001). Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. Dev. Biol. *234*, 275–288.

Blekhman, R., Man, O., Herrmann, L., Boyko, A.R., Indap, A., Kosiol, C., Bustamante, C.D., Teshima, K.M., and Przeworski, M. (2008). Natural selection on genes that underlie human disease susceptibility. Curr. Biol. *18*, 883–889.

Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S., and Van de Peer, Y. (2006). The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. *7*, R43.

Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., Karchin, R., Kinzler, K.W., Vogelstein, B., and Nowak, M.A. (2010). Accumulation of driver and passenger mutations during tumor progression. Proc. Natl. Acad. Sci. USA *107*, 18545–18550.

Brunet, F.G., Roest Crollius, H., Paris, M., Aury, J.M., Gibert, P., Jaillon, O., Laudet, V., and Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Mol. Biol. Evol. *23*, 1808–1816.

Cai, J.J., Borenstein, E., Chen, R., and Petrov, D.A. (2009). Similarly strong purifying selection acts on human disease genes of all evolutionary ages. Genome Biol. Evol. *1*, 131–144.

Ciocan, C.M., Moore, J.D., and Rotchell, J.M. (2006). The role of ras gene in the development of haemic neoplasia in Mytilus trossulus. Mar. Environ. Res. Suppl. *62*, S147–S150.

Dickerson, J.E., and Robertson, D.L. (2012). On the origins of Mendelian disease genes in man: the impact of gene duplication. Mol. Biol. Evol. *29*, 61–69.

Domazet-Loso, T., and Tautz, D. (2008). An ancient evolutionary origin of genes associated with human genetic diseases. Mol. Biol. Evol. *25*, 2699–2707.

Esteban, L.M., Vicario-Abejón, C., Fernández-Salguero, P., Fernández-Medarde, A., Swaminathan, N., Yienger, K., Lopez, E., Malumbres, M., McKay, R., Ward, J.M., et al. (2001). Targeted genomic disruption of H-ras and N-ras, individually or in combination, reveals the dispensability of both loci for mouse growth and development. Mol. Cell. Biol. *21*, 1444–1452.

Evangelisti, A.M., and Conant, G.C. (2010). Nonrandom survival of gene conversions among yeast ribosomal proteins duplicated through genome doubling. Genome Biol. Evol. *2*, 826–834.

Evlampiev, K., and Isambert, H. (2007). Modeling protein network evolution under genome duplication and domain shuffling. BMC Syst. Biol. *1*, 49.

Evlampiev, K., and Isambert, H. (2008). Conservation and topology of protein interaction networks under duplication-divergence evolution. Proc. Natl. Acad. Sci. USA *105*, 9863–9868.

Fernández, A., and Lynch, M. (2011). Non-adaptive origins of interactome complexity. Nature *474*, 502–505.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., et al. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. *39*(Database issue), D945–D950.

Freeling, M., and Thomas, B.C. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. *16*, 805–814.

Furney, S.J., Albà, M.M., and López-Bigas, N. (2006). Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. BMC Genomics *7*, 165.

Gibson, T.J., and Spring, J. (1998). Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. Trends Genet. *14*, 46–49, discussion 49–50.

Gout, J.F., Duret, L., and Kahn, D. (2009). Differential retention of metabolic genes following whole-genome duplication. Mol. Biol. Evol. *26*, 1067–1072.

Gout, J.F., Kahn, D., and Duret, L.; Paramecium Post-Genomics Consortium. (2010). The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet. *6*, e1000944.

Guan, Y., Dunham, M.J., and Troyanskaya, O.G. (2007). Functional analysis of gene duplications in Saccharomyces cerevisiae. Genetics *175*, 933–943.

Hakes, L., Pinney, J.W., Lovell, S.C., Oliver, S.G., and Robertson, D.L. (2007). All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol. *8*, R209.

Higgins, M.E., Claremont, M., Major, J.E., Sander, C., and Lash, A.E. (2007). CancerGenes: a gene selection resource for cancer genome projects. Nucleic Acids Res. *35*(Database issue), D721–D726.

Hughes, A.L. (1994). The evolution of functionally novel proteins after gene duplication. Proc. Biol. Sci. *256*, 119–124.

Huminiecki, L., and Heldin, C.H. (2010). 2R and remodeling of vertebrate signal transduction engine. BMC Biol. *8*, 146.

Ise, K., Nakamura, K., Nakao, K., Shimizu, S., Harada, H., Ichise, T., Miyoshi, J., Gondo, Y., Ishikawa, T., Aiba, A., and Katsuki, M. (2000). Targeted deletion of the H-ras gene decreases tumor formation in mouse skin carcinogenesis. Oncogene *19*, 2951–2956.

Kellis, M., Birren, B.W., and Lander, E.S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature *428*, 617–624.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2009). Human Protein Reference Database—2009 update. Nucleic Acids Res. *37*(Database issue), D767–D772.

Lin, Y.S., Hwang, J.K., and Li, W.H. (2007). Protein complexity, gene duplicability and gene dispensability in the yeast genome. Gene *387*, 109–117.

Lynch, M., and Force, A. (2000a). Gene duplication and the origin of interspecific genomic incompatibility. Am. Nat. *156*, 590–605.

Lynch, M., and Force, A. (2000b). The probability of duplicate gene preservation by subfunctionalization. Genetics *154*, 459–473.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. USA *102*, 5454–5459.

Makino, T., and McLysaght, A. (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc. Natl. Acad. Sci. USA *107*, 9270–9274.

Nadeau, J.H., and Sankoff, D. (1997). Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics *147*, 1259–1266.

Ohno, S. (1970). Evolution by Gene Duplication (New York: Springer-Verlag).

Papp, B., Pál, C., and Hurst, L.D. (2003). Dosage sensitivity and the evolution of gene families in yeast. Nature *424*, 194–197.

Pearl, J. (2001). Direct and indirect effects. Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence, San Francisco, CA: Morgan Kaufmann, 411–420.

Pearl, J. (2011). The Mediation Formula: A guide to the assessment of causal pathways in nonlinear models. In Causality: Statistical Perspectives and Applications, C. Berzuini, P. Dawid, and L. Bernardinelli, eds. (United Kingdom: John Wiley & Sons), pp. 151–175.

Pufall, M.A., and Graves, B.J. (2002). Autoinhibitory domains: modular effectors of cellular regulation. Annu. Rev. Cell Dev. Biol. *18*, 421–462.

Putnam, N.H., Butts, T., Ferrier, D.E., Furlong, R.F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J.K., et al. (2008). The amphioxus genome and the evolution of the chordate karyotype. Nature *453*, 1064–1071.

Qian, W., and Zhang, J. (2008). Gene dosage and gene duplicability. Genetics *179*, 2319–2324.

Robert, J. (2010). Comparative study of tumorigenesis and tumor immunity in invertebrates and nonmammalian vertebrates. Dev. Comp. Immunol. *34*, 915–925.

Robins, J.M., and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. Epidemiology *3*, 143–155.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. (2010). CORUM: the comprehensive resource of mammalian protein complexes–2009. Nucleic Acids Res. *38*, 497–501.

Sémon, M., and Wolfe, K.H. (2007). Consequences of genome duplication. Curr. Opin. Genet. Dev. *17*, 505–512.

Sémon, M., and Wolfe, K.H. (2008). Preferential subfunctionalization of slow-evolving genes after allopolyploidization in *Xenopus laevis*. Proc. Natl. Acad. Sci. USA *105*, 8333–8338.

Veitia, R.A. (2002). Exploring the etiology of haploinsufficiency. Bioessays *24*, 175–184.

Veitia, R.A. (2010). A generalized model of gene dosage and dominant negative effects in macromolecular complexes. FASEB J. *24*, 994–1002.

Wolfe, K. (2000). Robustness—it's not where you think it is. Nat. Genet. *25*, 3–4.

Zanivan, S., Cascone, I., Peyron, C., Molineris, I., Marchio, S., Caselle, M., and Bussolino, F. (2007). A new computational approach to analyze human protein complexes and predict novel protein interactions. Genome Biol. *8*, R256.

Zeevi, D., Sharon, E., Lotan-Pompan, M., Lubling, Y., Shipony, Z., Raveh-Sadka, T., Keren, L., Levo, M., Weinberger, A., and Segal, E. (2011). Compensation for differences in gene copy number among yeast ribosomal proteins is encoded within their promoters. Genome Res. *21*, 2114–2128.

Zhang, J., Feuk, L., Duggan, G.E., Khaja, R., and Scherer, S.W. (2006). Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet. Genome Res. *115*, 205–214.

la sensibilité des techniques d'imagerie, permettront de continuer à développer de nouveaux rapporteurs. En outre, les outils de détection et de suivi des cellules, ainsi que l'analyse des oscillations permettront de développer le même type d'approche chez d'autres vertébrés. ◊

A technical breakthrough for understanding segmentation clock dynamics and synchrony

LIENS D'INTÉRÊT

Les auteurs déclarent n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

RÉFÉRENCES

1. Pourquié O. Vertebrate segmentation: from cyclic gene networks to scoliosis. *Cell* 2011 ; 145 : 650–63.
2. Soroldoni D, Oates AC. Live transgenic reporters of the vertebrate embryo's segmentation clock. *Curr Opin Genet Dev* 2011 ; 21 : 600–5.
3. Masamizu Y, Ohtsuka T, Takashima, Y, *et al.* Real-time imaging of the somite segmentation clock: revelation of unstable oscillators in the individual presomitic mesoderm cells. *Proc Natl Acad Sci USA* 2006 ; 103 : 1313–8.
4. Aulehla A, Wiegraebe W, Baubet V, *et al.* A beta-catenin gradient links the clock and wavefront systems in mouse embryo segmentation. *Nat Cell Biol* 2008 ; 10 : 186–93.
5. Takashima Y, Ohtsuka T, Gonzalez A, *et al.* Intronic delay is essential for oscillatory expression in the segmentation clock. *Proc Natl Acad Sci USA* 2011 ; 108 : 3300–5.
6. Delaune EA, François P, Shih NP, Amacher SL. Single-cell-resolution imaging of the impact of Notch signaling and mitosis on segmentation clock dynamics. *Dev Cell* 2012 ; 23 : 995–1005.
7. Gajewski M, Sieger D, Alt B, *et al.* Anterior and posterior waves of cyclic her1 gene expression are differentially regulated in the presomitic mesoderm of zebrafish. *Development* 2003 ; 130 : 4269–78.
8. Jiang YJ, Aerne BL, Smithers L, *et al.* Notch signalling and the synchronization of the somite segmentation clock. *Nature* 2000 ; 408 : 475–9.
9. Horikawa K, Ishimatsu K, Yoshimoto E, *et al.* Noise-resistant and synchronized oscillation of the segmentation clock. *Nature* 2006 ; 441 : 719–23.

NOUVELLE

Évolution et cancer

Expansion des familles de gènes dangereux par duplication du génome

Séverine Affeldt[1*], Param Priya Singh[1*], Ilaria Cascone[2], Rasim Selimoglu[2], Jacques Camonis[2], Hervé Isambert[1]

[1]CNRS-UPMC UMR168, [2]Inserm U830, Institut Curie, Centre de recherche, 26, rue d'Ulm, 75248 Paris, France.
*Contribution égale des auteurs
herve.isambert@curie.fr

› Si la conservation des gènes essentiels à la vie des organismes se conçoit intuitivement bien, à l'inverse, l'étonnante expansion des familles de gènes à l'origine des cancers ou d'autres maladies génétiques chez les vertébrés pose question. Alors qu'on pourrait supposer que la multiplication de ces gènes « dangereux » confère malgré tout un avantage sélectif, nos travaux récents [1] suggèrent en fait que ces gènes ont été multipliés et conservés en raison de leur dangerosité à la suite de deux accidents génétiques majeurs correspondant à des duplications globales de génome.

De l'expansion des familles de gènes dangereux chez les vertébrés

Pour comprendre l'origine de l'expansion des familles de gènes dangereux chez les vertébrés, il faut remonter à l'ancêtre commun de tous les vertébrés, il y a quelque 500 millions d'années. Par un mécanisme presque toujours létal, mais qui a joué un rôle essentiel au cours de l'évolution, notre lignée encore invertébrée a entièrement dupliqué son génome deux fois de suite et a survécu à ces deux accidents génétiques majeurs. Ces deux duplications globales du génome ont conduit à l'émergence et à la complexification des vertébrés dont certains gènes ont conservés jusqu'à quatre copies. Au total, un quart à un tiers de nos gènes serait directement issu de ces deux duplications de génome à l'origine des vertébrés [2]. Ces gènes sont appelés gènes ohnologues en l'honneur du généticien Susumu Ohno qui fut le premier à avancer l'hypothèse de ces duplications globales du génome chez les vertébrés [3].

Après une duplication globale du génome, les organismes perdent généralement près de 80 à 90 % des gènes dupliqués, ce qui entraîne une expansion hétérogène de leurs voies de signalisation *(Figure 1)* et de leurs réseaux de gènes [4-6]. Cependant, de façon surprenante, on constate que les copies ohnologues retenues dans le génome humain comportent un nombre élevé de gènes dangereux, c'est-à-dire présentant une forte susceptibilité aux mutations délétères dominantes comme les oncogènes notamment. Certains de ces gènes dangereux ont même gardé leurs quatre copies depuis l'origine des vertébrés ! C'est le cas par exemple des gènes *RalGEF* (*Ral guanine-nucleotide exchange factor*) *(Figure 1)* qui activent les voies Ras-Ral impliquées dans la migration et la prolifération cellulaires dans des tumeurs [7]. De même, le gène *Ras*, qu'on retrouve en un seul exemplaire chez les invertébrés comme la drosophile *(Figure 1)*, a conservé chez la plupart des vertébrés trois ohnologues proto-oncogéniques (*KRas*, *HRas* et *NRas*) *(Figure 1)* qui présentent des mutations constitutivement actives dans plus de 25 % des cas de cancer chez l'homme.

DOI : 10.1051/medsci/2013294008

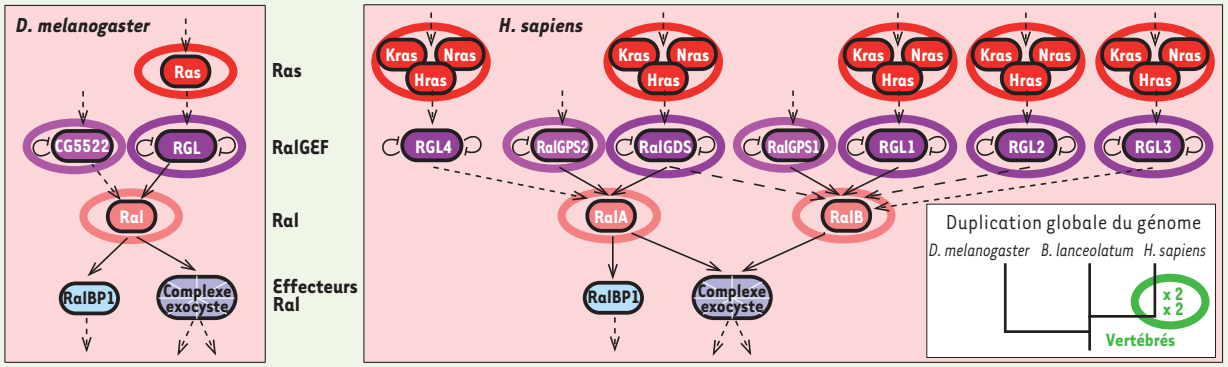


Figure 1. ***Expansion des voies de signalisation Ras-Ral.*** Voies de signalisation Ras-Ral chez la drosophile (***A***) et chez l'homme (***B***). Après les deux duplications globales de génomes chez l'ancêtre des vertébrés, chez l'homme, le gène *RGL* a conservé ses quatre copies ohnologues (*RalGDS*, *RGL1*, *RGL2*, *RGL3*), le gène *Ras* a conservé trois copies ohnologues (*Kras*, *Hras*, *Nras*) et le gène *Ral* a conservé deux copies ohnologues (*RalA*, *RalB*) [7]. Cet exemple illustre la nécessaire réorganisation des voies de signalisation et des réseaux d'interactions protéine-protéine après un évènement de duplication globale du génome, et l'élimination par divergence d'une partie des gènes dupliqués [4-6].

Au delà de ces exemples, nous avons effectué une étude d'exploration de données à grande échelle sur l'ensemble des gènes humains, à partir des bases de données accessibles en ligne (comme COSMIC [8], CancerGenes [9]) et de publications. Ceci nous a permis de mettre en évidence une forte association entre la dangerosité des ohnologues et leur rétention dans le génome humain. Dans ces analyses, nous avons considéré différentes classes de gènes susceptibles d'être affectés par des mutations délétères. Ces classes comprennent notamment des gènes dont l'implication dans les cancers est connue et des gènes dont les mutations induisent typiquement des phénotypes délétères dominants, par exemple *via* une perte d'interaction d'auto-inhibition entraînant un gain de fonction permanent pour le gène mutant.

Les résultats obtenus [1] indiquent que les 8 095 gènes impliqués dans des cancers ou dans des maladies génétiques comportent significativement plus d'ohnologues que l'ensemble des 20 506 gènes codant pour les protéines humaines, soit 48 % contre 35 % (48 % ; 3 844/8 095 ; $p = 1{,}3 \times 10^{-129}$, test χ^2). En outre, ce biais de rétention augmente fortement dans le cas des oncogènes (61 % ; 493/813 ; $p = 1{,}4 \times 10^{-54}$, test χ^2) ou des gènes dont la protéine est auto-inhibée (76 % ; 350/461 ; $p = 2{,}7 \times 10^{-77}$, test χ^2). De même, lorsque l'on considère des classes de gènes combinant plusieurs facteurs de dangerosité, telles que la classe des oncogènes dont la protéine est auto-inhibée, ces biais de rétention dépassent les 90 % (91 % ; 104/114 ; $p = 6{,}9 \times 10^{-37}$, test χ^2). À l'inverse, nous avons montré que les gènes associés à des mutations délétères récessives chez l'homme, ainsi que la plupart des gènes essentiels connus chez la souris, n'ont pas conservé un excès d'ohnologues. Il apparaît donc que la rétention des gènes ohnologues serait directement liée à leur susceptibilité aux mutations délétères dominantes et non pas à leur nature fondamentale pour l'organisme. Nous avons en fait démontré [1], au moyen d'analyses statistiques d'inférences bayésiennes, telles que l'analyse de Médiation [10], que la dangerosité est bien la cause principale de la rétention des gènes ohnologues chez l'homme. En particulier, les équilibres entre les niveaux d'expression génétique, fréquemment proposés comme la principale cause des biais de rétention des ohnologues, semblent en fait résulter indirectement des effets des mutations délétères dominantes [1].

Le mécanisme de conservation des ohnologues

Pourquoi les multiples copies ohnologues de ces gènes dangereux, à l'origine de nombreux cancers et maladies génétiques sévères, n'ont-elles pas été éliminées chez les vertébrés ? Pour le comprendre, il faut avoir en tête deux points essentiels *(Figure 2)* : (1) la duplication globale du génome, lorsqu'elle n'est pas létale, implique nécessairement l'apparition d'une nouvelle espèce composée d'individus possèdant tous initialement leurs gènes en double ; et (2) l'inégalité des gènes face aux mutations. Alors que la plupart des gènes tendent à perdre leur fonction par mutation, les gènes dangereux se caractérisent par le fait que leurs mutations entraînent fréquemment une suractivation, c'est-à-dire un gain de fonction, plutôt qu'une perte de fonction. En général, la perte de fonction d'un ohnologue ne pose pas de problème tant qu'il reste une copie fonctionnelle de ce gène, ce qui conduit à l'élimination progressive d'une des copies de la plupart des ohnologues non dangereux. En revanche, la survenue de mutations conduisant à des gains de fonction ou, plus généralement, à des phénotypes dominants délétères qui caractérisent les ohnologues dangereux, va entraîner des pathologies du développement ou

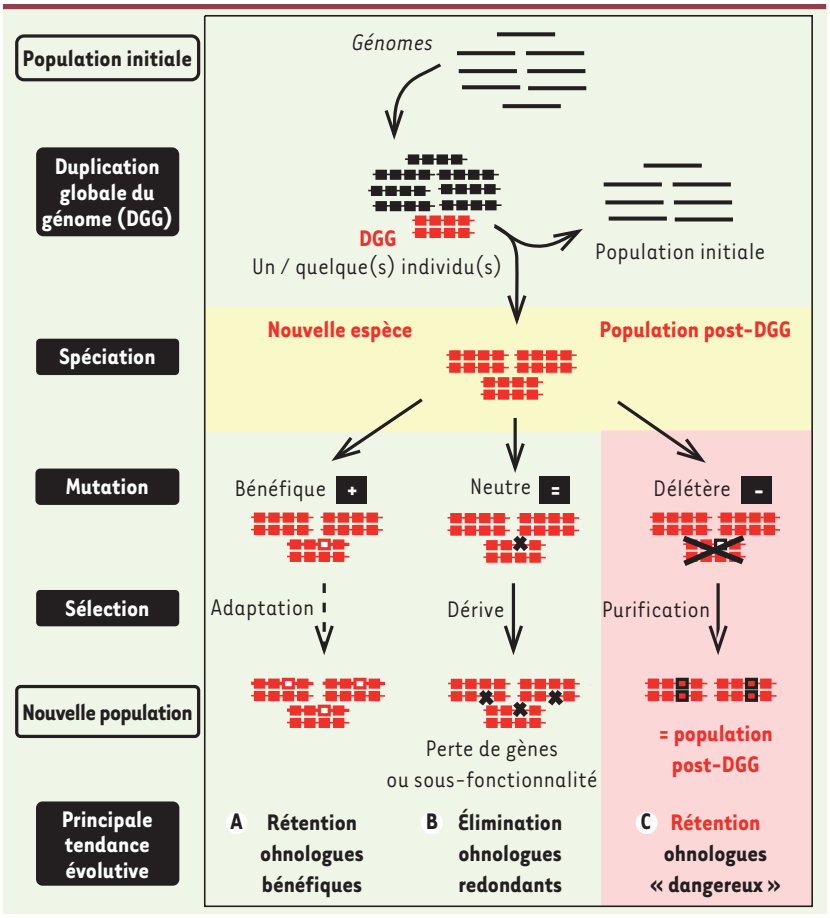


Figure 2. ***Évolution des gènes dupliqués après duplicaton globale du génome. A-C.*** Les lignes horizontales représentent le génome de différents individus. Les carrés symbolisent les gènes dupliqués (rouges) ou non dupliqués (noirs). Les croix noires représentent la perte d'un seul gène (petites croix) ou l'élimination d'un individu (grande croix), tandis que les carrés avec bordures indiquent les copies mutées. Les scénarios d'évolution sont décrits comme il se doit au niveau d'une population d'individus. Lorsqu'elle n'est pas létale, une duplication globale de génome chez un ou plusieurs individus de la population initiale (deux lignes horizontales rouges pour un individu avec duplication complète) implique nécessairement l'émergence d'une nouvelle espèce dont tous les gènes ont été dupliqués. La sélection de purification favorise alors indirectement la conservation des ohnologues dangereux non mutés dans le génome des individus qui survivent. Adapté de [1].

des tumeurs. Celles-ci pénalisent les organismes atteints et, plus ou moins directement, leur descendance qui finira par s'interrompre. Pour autant, les gènes dangereux impliqués ne sont pas éliminés mais au contraire conservés dans la population, puisqu'ils sont encore présents sous une forme non délétère dans le reste de la population issue de la duplication du génome *(Figure 2)*. Ce processus évolutif par élimination de mutants (sélection de purification) se distingue du concept d'avantage sélectif généralement associé à l'évolution (sélection naturelle ou adaptative).

La multiplication des gènes dangereux chez les vertébrés est donc liée à ce phénomène spécifique et très rare de duplication globale du génome qui existe en fait dans la plupart des branches eucaryotes. Sa compréhension doit se faire au niveau de la génétique des populations au sein des nouvelles espèces issues d'une duplication de génome. Tout se passe comme si les individus de ces populations n'avaient pas pu se débarrasser de nombreux gènes dangereux redondants, directement fixés par spéciation dans leur génome dupliqué et non pas fixés progressivement grâce à un avantage sélectif comme pour la plupart des gènes dupliqués individuellement. Ensuite, les gènes ohnologues conservés se différencient et deviennent souvent des acteurs majeurs du développement, de la signalisation et de la régulation cellulaires. Par exemple, les cadhérines, sorte de colle qui lient les cellules entre elles, ont conservé de multiples ohnologues exprimés dans différents tissus, comme la E-cadhérine qui lie les cellules épithéliales entre elles, ou la N-cadhérine exprimée dans le tube neural et les neurones. Mais les mutations des cadhérines qui entraînent la décohésion des cellules entre elles sont aussi impliquées dans la migration des cellules tumorales et leur dissémination vers d'autres organes.

Conclusions

Ces accidents génétiques majeurs de doublement du génome survenus il y a 500 millions d'années dans l'évolution des vertébrés ont donc permis l'émergence d'organismes plus complexes, mais aussi la multiplication des gènes dangereux chez les vertébrés. Ces résultats éclairent d'un point de vue évolutif nouveau l'expansion des familles de gènes fréquemment impliquées dans des maladies génétiques et de nombreux cancers. ◊

Evolution and cancer: expansion of dangerous gene repertoire by whole genome duplications

LIENS D'INTÉRÊT

Les auteurs déclarent n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

RÉFÉRENCES

1. Singh PP, Affeldt S, Cascone I, *et al*. On the expansion of dangerous gene repertoires by whole-genome duplications in early vertebrates. *Cell Rep* 2012 ; 2 : 1387-98.
2. Makino T, McLysaght A. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci USA* 2010 ; 107 : 9270-74.

3. Ohno S. *Evolution by gene duplication*. New York : Springer-Verlag, 1970.
4. Evlampiev K, Isambert H. Modeling protein network evolution under genome duplication and domain shuffling. *BMC Syst Biol* 2007 ; 1 : 49.
5. Evlampiev K, Isambert H. Conservation and topology of protein interaction networks under duplication-divergence evolution. *Proc Natl Acad Sci USA* 2008 ; 105 : 9863-8.
6. Stein RR, Isambert H. Logistic map analysis of biomolecular network evolution. *Phys Rev E* 2011 ; 84 : 051904.
7. Cascone I, Selimoglu R, Ozdemir C, *et al.* Distinct roles of RalA and RalB in the progression of cytokinesis are supported by distinct RalGEFs. *EMBO J* 2008 ; 27 : 2375-87.
8. Forbes SA, Bindal N, Bamford S, *et al.* COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res* 2011 ; 39 : D945–D950.
9. Higgins ME, Claremont M, Major JE, *et al.* CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Res* 2007 ; 35 : D721–D726.
10. Pearl J. *Causality: models, reasoning and inference*. New York : Cambridge University Press, 2009.

NOUVELLE

Aldéhydes et anémie de Fanconi

L'ennemi de l'intérieur

Frédéric P.M. Langevin, Juan I. Garaycoechea, Gerry P. Crossan, Ketan J. Patel

Medical Research Council, Laboratory of Molecular Biology, Francis Crick avenue, Cambridge Biomedical Campus, Cambridge, CB2 0QH, Royaume-Uni.
lf95@mrc-lmb.cam.ac.uk
kjp@mrc-lmb.cam.ac.uk

› La survie d'un organisme multicellulaire est étroitement liée au maintien de l'homéostasie dans l'ensemble des tissus constituant cet organisme. Cette capacité à maintenir cet équilibre vital est assurée par des populations de cellules souches, *via* trois caractéristiques principales : (1) la possibilité de se diviser et de renouveler ce compartiment en générant de nouvelles cellules souches ; (2) l'absence de fonctions tissulaires spécifiques ; et (3) la capacité de se différencier en cellules spécialisées. La moelle osseuse, où résident les cellules souches hématopoïétiques (CSH) responsables de la production des cellules du sang, constitue un tissu modèle pour l'étude des cellules souches somatiques. En effet, l'identification de multiples marqueurs de surface et le développement d'anticorps dirigés contre ces marqueurs et utilisables en cytométrie de flux (FACS, *fluorescent activated cell sorter*), permettent de discriminer et d'isoler les divers précurseurs des cellules sanguines.

Les cellules souches sont sensibles, comme toute cellule, aux conditions de stress, en particulier aux dommages de l'ADN. Ainsi, au cours du temps, l'accumulation de mutations contribue au déclin des capacités des cellules souches. L'existence de mécanismes de réparation de l'ADN est donc essentielle au maintien des populations de cellules souches [1]. Une étude récente de notre groupe décrit le rôle d'un système de réparation de l'ADN et la nature génotoxique de métabolites endogènes dans les CSH de souris [2].

L'anémie de Fanconi, une pathologie s'accompagnant d'un défaut de réparation de l'ADN

La thématique de recherche principale de notre équipe concerne l'étude des mécanismes de réparation de l'ADN et, en particulier, celle des gènes impliqués dans l'anémie de Fanconi. Notre approche est essentiellement génétique et se base sur des modèles cellulaires et murins (souris génétiquement modifiées ou *knockout* [KO]). Du nom du pédiatre suisse Guido Fanconi qui décrivit cette maladie génétique rare au début du XX^{e} siècle [3], l'anémie de Fanconi se manifeste dès l'enfance par des malformations squelettiques, une forte prédisposition à certains types de cancers (en particulier des leucémies myéloïdes) et une aplasie médullaire progressive. Les cellules de ces patients sont très sensibles à des drogues anticancéreuses, telles que la cisplatine ou la mitomycine C. Cette sensibilité constitue la base du test de diagnostic dit de « cassures chromosomiques » [4]. En effet, en liant de façon covalente deux bases opposées (interbrins d'ADN) ou adjacentes (intrabrin d'ADN), ces drogues induisent des lésions (ponts inter- ou intrabrins, *DNA crosslinks*) de l'ADN qui bloquent la réplication et la transcription. Ces ponts interbrins ont pour conséquence une augmentation des cassures chromosomiques. Dans les cellules de patients atteints d'anémie de Fanconi, ceci est dû à un défaut de réparation de l'ADN. À ce jour, une quinzaine de gènes ont été identifiés, codant pour des protéines impliquées dans cette voie de réparation des lésions pontantes de l'ADN.

Les aldéhydes, source de dommages à l'ADN

Les cassures chromosomiques se manifestent spontanément dans les cellules de patients, en dehors de la présence de drogues comme la cisplatine, ce qui suggère l'existence d'agents génotoxiques présents de façon naturelle dans les cellules de ces patients. Notre groupe a confirmé cette hypothèse en montrant en 2011 que les aldéhydes produits par le métabolisme cellulaire (en particulier l'acétaldéhyde) étaient à l'origine de dommages à l'ADN dans les cellules de patients atteints d'anémie de Fanconi [5]. Des souris double KO pour les

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Formal Comment

Human Dominant Disease Genes Are Enriched in Paralogs Originating from Whole Genome Duplication

Param Priya Singh¤, Séverine Affeldt, Giulia Malaguti, Hervé Isambert*

CNRS-UMR168, UPMC, Institut Curie, Research Center, Paris, France

PLOS Computational Biology recently published an article by Chen, Zhao, van Noort, and Bork [1] reporting that, in contrast to duplicated nondisease genes, human monogenic disease (MD) genes are (1) enriched in duplicates (in agreement with earlier reports [2–5]) and (2) more functionally similar to their closest paralogs based on sequence conservation and expression profile similarity. Chen et al. then proposed that human MD genes frequently have functionally redundant paralogs that can mask the phenotypic effects of deleterious mutations.

We would like to point out here two lines of evidence that appear more relevant to the explanation of this surprising enrichment of human disease genes in duplicates. The first line of evidence indicates that human gene duplicates should be distinguished depending on whether they originate from small-scale duplication (SSD) or from the two rounds of whole genome duplication (WGD) that occurred in early vertebrates some 500 million years ago. In fact, as shown quantitatively below using Chen et al.'s dataset, human MD genes are actually depleted, not enriched, in SSD duplicates, whereas they are clearly enriched in WGD duplicates when compared to nondisease genes. This opposite retention pattern cannot be explained by a selection mechanism independent of the SSD or WGD origin of MD gene duplicates. The second line of evidence concerns the mode of inheritance of human MDs, which provides a more stringent criterion than sequence conservation or coexpression profile to assess the likelihood of functional compensation by paralogs of MD genes. In particular, the recessiveness of a human disease is expected to be a prerequisite for functional compensation by a paralog gene. Indeed, autosomal dominant MDs are unlikely to experience significant functional compensation from a different locus, since even a perfectly functional allele is unable to mask the deleterious phenotypic effects of a dominant allelic mutant on the same heterozygote locus.

We first address the difference between SSD duplicates and WGD duplicates, also called "ohnologs" after Susumu Ohno's early "2R hypothesis" [6], which has now been firmly established [7]. The importance of distinguishing between SSD and WGD duplicates in the human genome has already been reported in a number of papers [2–4,8], including our own [5,9]. As shown in Figure 1A, human genes tend to partition into three main gene categories with respect to duplicates: those with WGD but no SSD duplicates (about 28%), those with SSD but no WGD duplicates (about 41%), and singletons without WGD or SSD duplicates (about 24%), while human genes with both WGD and SSD duplicates are relatively rare (about 7%). Gene families enriched either in WGD or SSD duplicates also correspond to distinct functional classes [2,8], with WGD genes frequently involved in signaling, regulation, and development, whereas SSD genes are typically implicated in different functions such as antigen processing, immune response, and metabolism.

In addition, human disease genes have been shown to be significantly enriched in WGD duplicates, while they are rather depleted in SSD duplicates [2,5,8,9]. This could not be seen with Chen et al.'s dataset, which lumps together all gene duplicates irrespective of their WGD or SSD origin. In fact, using the same monogenic disease (MD) dataset, we could readily extend these earlier results, as depicted in Figure 1B. MD genes are significantly enriched in ohnologs, 38.3% versus 27.7% ($p = 1.58 \times 10^{-25}$; Fisher's Exact [FE] test), while showing at the same time a significant depletion in both singletons, 16.5% versus 23.7% ($p = 7.67 \times 10^{-20}$; FE test), and SSD, 36.1% versus 41.6% ($p = 2.75 \times 10^{-6}$; FE test). MD genes are more specifically depleted in recent SSD, 9.2% versus 17.3% ($p = 4.1 \times 10^{-50}$; FE test), while WGD-old and older SSD of MD genes are not significantly biased, i.e., 9.9% versus 9.2% ($p = 0.12$; FE test) and 17% versus 15.5% ($p = 0.001$; FE test), respectively (see below). These results demonstrate that, although MD genes retain significantly more duplicates than singletons (Figure 1B), these duplicates are primarily enriched in ohnologs and not SSD copies, as compared to the relative WGD and SSD content of the entire human genome (Figure 1A, Dataset S1).

To explain the global enrichment in MD gene duplicates, Chen et al. noticed that coexpressions between MDs and their closest paralogs are in general higher than that of nondisease genes ($p = 0.00298$, Figure 2B in [1]), which they interpret as evidence that "functional compensation by duplication of genes masks the phenotypic effects of deleterious mutations and reduces the probability of purging the defective genes from the human population." In particular, the retention of MD gene duplicates should be favored by the higher functional redundancy of recent, less-diverged duplicates. However, investigating the age of SSD duplicates from MD genes suggests rather the opposite, as MD genes tend to have fewer recent SSD than old SSD duplicates, as compared to nondisease (ND) genes (Figures 1A and B). In particular, focusing on genes with SSD but no ohnolog, we found that 9.2% [respectively 17%] of MD genes have SSD that are more recent [respectively ancient] than the two rounds of whole-

Citation: Singh PP, Affeldt S, Malaguti G, Isambert H (2014) Human Dominant Disease Genes Are Enriched in Paralogs Originating from Whole Genome Duplication. PLoS Comput Biol 10(7): e1003754. doi:10.1371/journal.pcbi.1003754

Editor: Alon Keinan, Cornell University, United States of America

Published July 31, 2014

Funding: PPS acknowledges a PhD fellowship from Erasmus Mundus (Université Pierre et Marie Curie) and La Ligue Contre Le Cancer; SA and GM acknowledge a PhD fellowship from Ministry of Higher Education and Research. HI acknowledges funding from Foundation Pierre-Gilles de Gennes. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: herve.isambert@curie.fr

¤ Current address: Department of Genetics, Stanford University, Stanford, California, United States of America

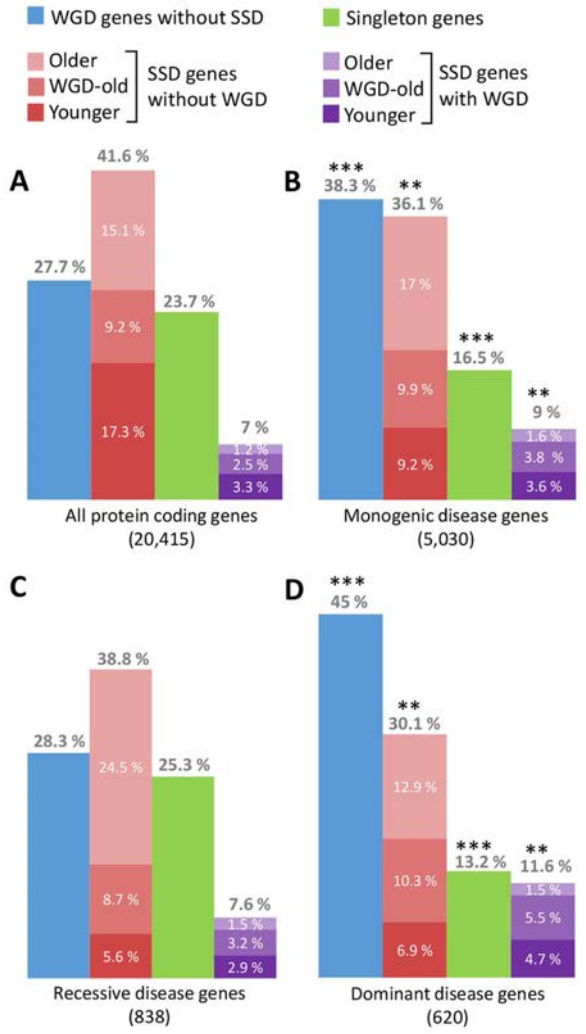


Figure 1. Distributions of WGD, SSD, and singletons in (A) the whole human genome, (B) monogenic disease (MD) genes [1]**, (C) recessive MD genes, and (D) dominant MD genes.** (***) corresponds to highly significant deviations ($p<10^{-6}$, FE test) and (**) to significant deviations ($p<10^{-3}$, FE test) from the references in (A). Note that recessive MD genes (C) do not show any significant deviations in WGD, SSD, or singleton contents ($p>0.3$, FE test), although taking into account the age of SSD duplicates reveals a relative lack of recent SSD genes in MD genes (see text).
doi:10.1371/journal.pcbi.1003754.g001

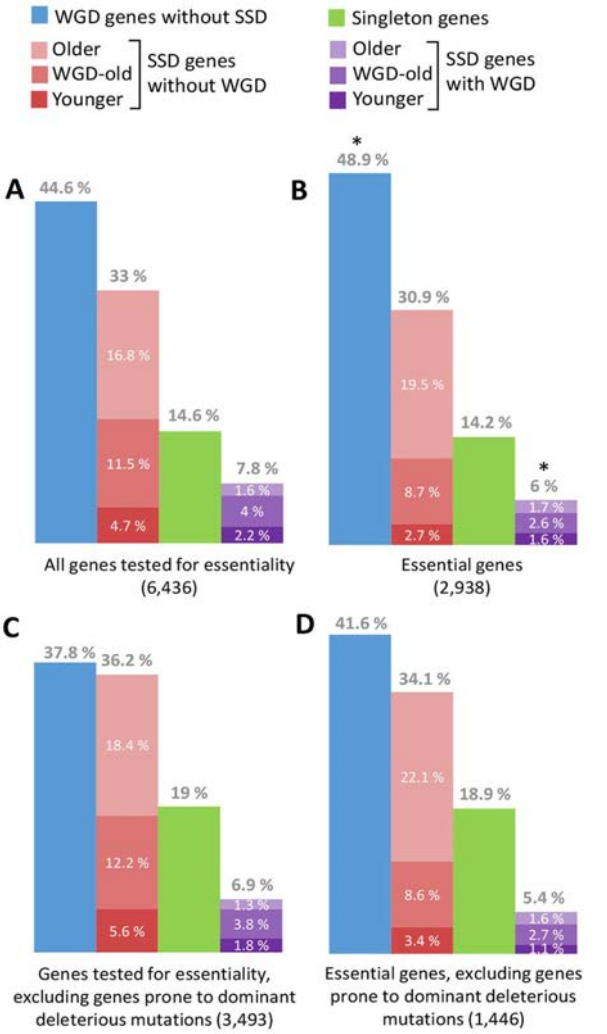


Figure 2. Distributions of WGD, SSD, and singletons for human orthologs of mouse genes (A) tested for essentiality in mouse [13]**, (B) found to be essential in mouse, and (C and D) after removing dominant disease genes, oncogenes, and genes with dominant negative mutations or autoinhibitory folds** [5]**.** (*) corresponds to small deviations ($10^{-3}<p<0.05$, FE test) from the references in (A). Note that human orthologs of essential genes in mouse do not show any significant deviations in WGD, SSD, or singleton contents ($p>0.05$, FE test) once dominant disease genes, oncogenes, and genes with dominant negative mutations or auto-inhibitory folds have been removed. Yet, taking into account the age of SSD duplicates reveals a relative lack of recent SSD genes in essential genes (see text).
doi:10.1371/journal.pcbi.1003754.g002

genome duplication, while the overall genome exhibits 17.3% [respectively 15.1%] instead ($p=4.5\times10^{-34}$; FE test). This suggests that the functional compensation, which can occur between functionally redundant duplicates, leads to a depletion (not an enrichment) of MD genes with recent SSD, in agreement with an earlier report [10].

In addition, we note that, while recent gene duplicates might be able to mask the phenotypic effect of recessive (e.g., loss-of-function) mutations, dominant (e.g., gain-of-function or dominant negative) mutations should typically lead to deleterious phenotypic effects regardless of the presence of any functionally redundant paralog at a different locus on the human genome.

In order to assess the extent of possible functional compensation on the retention of MD gene duplicates, we have thus investigated the mode of inheritance of human MDs. To this end, we retrieved the available information on the dominance and recessiveness of MDs from Online Mendelian Inheritance in Man (OMIM) [11] and Blekhman et al. [12]. Manual curation yielded 620 autosomal dominant and 838 autosomal recessive MD genes after excluding sex-linked genes and MD genes documented as both dominant and recessive (Dataset S1).

Using Chen et al.'s dataset and analysis, we then found that autosomal recessive MD gene duplicates (with possible functional compensation) do not exhibit significantly more correlated expression profiles than ND genes ($p=0.426$, Wilcoxon Rank

Sum Test, as compared to $p = 0.00298$ for all MD genes in Figure 2B in [1]), whereas autosomal dominant MD gene duplicates (with unlikely functional compensation) in fact exhibit significant expression profile correlations ($p = 0.00028$).

Moreover, looking for duplication biases of recessive versus dominant MDs confirmed that recessive MDs, which could in principle provide functional compensation, have not retained significantly more duplicates. Indeed, Figure 1C shows that recessive MDs do not present any biased retention of ohnologs, 28.3% versus 27.7% ($p = 0.79$; FE test); SSD duplicates, 38.8% versus 41.6% ($p = 0.31$; FE test); or singletons, 25.3% versus 23.7% ($p = 0.42$; FE test), as compared to their respective prevalence in the entire human genome (Figure 1A). These observations clearly show that the maintenance of recessive MD genes is largely independent of their WGD, SSD, or singleton status, suggesting limited effects of functional compensation by paralogs on the retention of gene duplicates associated to recessive MDs in human. By contrast, we observed (Figure 1D) that dominant MDs exhibit a strong enrichment in ohnologs, 45% versus 27.7% ($p = 1.8 \times 10^{-10}$; FE test), with concomitant depletions in both SSD, 30.1% versus 41.6% ($p = 0.0001$; FE test), and singletons, 13.2% versus 23.7% ($p = 1.59 \times 10^{-7}$; FE test). The same trend is observed for haploinsufficient and dominant negative genes [5]. This is unlikely to result from a functional compensation by paralogs because of the molecular genetics of dominance, as discussed above.

Finally, we investigated the enrichment in WGD and SSD duplicates of essential genes for which functional compensation could in principle be advantageous owing to the lethality of their double mutants. However, we found that human orthologs of mouse genes, reported as being "essential" genes from large-scale null mutant studies in mouse [13], are only slightly enriched in ohnologs, 48.9% versus 44.6% ($p = 0.02$, FE test), and hardly depleted in SSD, 30.9% versus 33% ($p = 0.14$, FE test), in which 44.6% and 33% are, respectively, the global proportions of ohnologs and SSD among the 6,436 genes tested for null mutation in mouse (Figures 2A and 2B). In fact, these small deviations, consistent with earlier findings [14], even become nonsignificant once genes with dominant allelic mutants are removed from the list of 6,436 genes tested for essentiality in mouse (Figures 2C and 2D), i.e., 41.6% versus 37.8% for ohnologs ($p = 0.1$, FE test) and 34.1% versus 36.2% for SSD ($p = 0.34$, FE test), in which 37.8% and 36.2% are, respectively, the global proportions of ohnologs and SSD among the 3,493 genes tested for null mutation in mouse after removing dominant disease genes, oncogenes, and genes with dominant negative mutations or autoinhibitory folds (Figures 2C and 2D) [5]. Hence, we could not find any significant enrichment in duplicates in support of possible functional compensation for essential genes, in broad agreement with early reports [15,16]. Moreover, taking into account the age distribution of SSD duplicates (including corrections for the visible age bias of genes tested for essentiality in mouse [Figures 1A and 2A]) [17,18], we actually found a relative lack of recent SSD of essential genes (Figure 2B and 2D) as observed for MD genes (Figure 1B–1D). This suggests evidence against functional redundancy in vertebrate essential genes, in agreement with an earlier report [17] and similar observations in yeast [19] and nematodes [20].

So, what could be the evolutionary mechanism behind the enhanced retention of WGD duplicates and relative depletion of SSD duplicates and singletons associated to MDs in humans (Figure 1B)? In other works [5,9], we proposed a population genetics model based on the observation that a major difference between SSD and WGD scenarios concerns the timing of fixation of gene duplicates. It is well-known that the SSD scenario starts with a gene duplication in the genome of a single individual, which subsequently needs to spread through the entire population to reach fixation. By contrast, the WGD scenario entails an initial fixation of duplicated gene pairs in the genome of all individuals in the small population, arising through WGD. This is because WGD typically induces a speciation event due to the ploidy incompatibility of the post-WGD individuals with the rest of the pre-WGD population. This population genetics model [9] for the fixation of SSD versus WGD duplicates then predicts that the enhanced retention of "dangerous" ohnologs prone to dominant deleterious mutations (as depicted in Figure 1D) is a direct consequence of purifying selection in post-WGD population, as most surviving individuals retain (nondeleterious) functional copies of their ohnologs that are prone to dominant deleterious mutations. By contrast, ohnologs prone to recessive deleterious mutations are more readily eliminated through loss-of-function mutations and are not expected to exhibit significant ohnolog retention bias (in agreement with Figure 1C). As for SSD duplicates, they are expected to be retained either from adaptive selection in large populations ($N > 10^5$) or from purifying selection in small populations ($N < 10^4$), in which SSD duplicates typically reach fixation by drift before their mutations actually occur, hence resembling the WGD scenario with an initial fixation of ohnologs through speciation in that case. This leads in principle to a complex retention pattern of SSD duplicates across evolutionary ages, in particular around WGD-induced population bottlenecks. Yet, overall it appears that genes with recent SSD duplicates are less likely to be MD genes than genes with WGD-old or older SSD duplicates (Figure 1).

All in all, we found that MD genes have preferentially retained WGD rather than SSD duplicates, as compared to nondisease genes. Yet, only dominant MD genes exhibit a clear enrichment in WGD duplicates, while the retention of duplicates of recessive MDs or essential genes, which might in principle experience functional compensation from paralogs, is in fact largely independent of their WGD, SSD, or singleton status. These results cannot be explained by the functional compensation hypothesis proposed in Chen et al. [1]. They are, however, consistent with a population genetics model taking into account the initial fixation of ohnologs through WGD-induced speciation and the ensuing purifying selection in post-WGD populations [5,9].

Materials and Methods

We obtained 20,415 protein coding genes in the human genome from Ensembl version 70 (Dataset S1). Ohnologs (7,075 genes) were obtained from [4], and SSDs (9,916 genes) were obtained by running an all-against-all BLASTp using the human proteins (see [5] for details). The genes which could not be classified as ohnologs or SSDs were taken to be the singleton genes (4,846 genes). The duplication timing of SSD genes was obtained from Ensembl compara [21] using BioMart.

MD genes were taken from Chen et al. [1]. We could map 5,030 of 5,134 MD genes on our dataset using BioMart. Inheritance status of the MD genes were obtained either from the inheritance section from OMIM entries [11] or from Blekhman et al. [12]. After careful manual curation, we could obtain 1,458 MD genes in which the inheritance pattern was unambiguously described as either autosomal dominant (620 genes) or autosomal recessive (838 genes).

Expression profile correlation between autosomal recessive genes and ND genes, and autosomal dominant genes and ND genes was performed using the R scripts provided by Chen et al. [1].

We obtained 6,436 mouse genes tested for null/knock-out mutations from the Mouse Genome Database (MGD) [13], as described in [5]. 2,938 of these 6,436 genes had lethal or infertile phenotypes and were classified as "essential." Human one-to-one orthologs of these genes were obtained using Ensembl BioMart. We also investigated the enrichment of essential genes in duplicates after removing dominant disease genes. To this end we considered multiple classes of genes susceptible to dominant mutations, including 620 dominant disease genes (from this report), 5,996 oncogenes [22], 566 dominant negative genes, and 461 genes having autoinhibitory folds [5].

Supporting Information

Dataset S1 Dataset used for the analysis. Gene IDs and symbols are from Ensembl v70. Values in the columns correspond to one of the following descriptions: N = no; Y = yes; Onc = oncogene; TS = tumor suppressor; O = others; AD = autosomal dominant; AR = autosomal recessive; XL = X-linked; and YL = Y-linked.
(XLSX)

Author Contributions

Conceived and designed the experiments: PPS HI. Performed the experiments: PPS SA GM HI. Analyzed the data: PPS SA GM HI. Wrote the paper: PPS SA HI.

References

1. Chen WH, Zhao XM, van Noort V, Bork P (2013) Human Monogenic Disease Genes Have Frequently Functionally Redundant Paralogs. PLoS Comput Biol 9: e1003073.
2. Makino T, McLysaght A (2010) Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc Natl Acad Sci USA 107: 9270–9274.
3. Dickerson JE, Robertson DL (2012) On the origins of Mendelian disease genes in man: the impact of gene duplication. Mol Biol Evol 29: 61–69.
4. Tinti M, Johnson C, Toth R, Ferrier D, MacKintosh C (2012) Evolution of signal multiplexing by 14-3-3-binding 2R-ohnologue protein families in the vertebrates. Open Biol 2: 120103.
5. Singh PP, Affeldt S, Cascone I, Selimoglu R, Camonis J, et al. (2012) On the expansion of "dangerous" gene repertoires by whole-genome duplications in early vertebrates. Cell Rep 2: 1387–1398.
6. Ohno S, Wolf U, Atkin NB (1968) Evolution from fish to mammals by gene duplication. Hereditas 59: 169–187
7. Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453: 1064–1071
8. Huminiecki L, Heldin CH (2010) 2R and remodeling of vertebrate signal transduction engine. BMC Biol 13: 8–146.
9. Malaguti G, Singh PP, Isambert H (2014) On the retention of gene duplicates prone to dominant deleterious mutations. Theor Popul Biol 93: 38–51.
10. Hsiao TL, Vitkup D (2008) Role of duplicate genes in robustness against deleterious human mutations. PLoS Genet 4: e1000014
11. McKusick VA (2007) Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet 80: 588–604.
12. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, et al. (2008) Natural selection on genes that underlie human disease susceptibility. Curr Biol 18: 883–889.
13. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, et al. (2012) The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. Nucleic Acids Res 40: D881–D886.
14. Makino T, Hokamp K, McLysaght A (2009) The complex relationship of gene duplication and essentiality. Trends Genet 25: 152–155.
15. Liao BY, Zhang J (2007) Mouse duplicate genes are as essential as singletons. Trends Genet 23: 378–381.
16. Liang H, Li WH (2007) Gene essentiality, gene duplicability and protein connectivity in human and mouse. Trends Genet 23: 375–378.
17. Su Z, Gu X (2008) Predicting the proportion of essential genes in mouse duplicates based on biased mouse knockout genes. J Mol Evol 67: 705–709.
18. Chen WH, Trachana K, Lercher MJ, Bork P (2012) Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. Mol Biol Evol 29: 1703–1706
19. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, et al. (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421: 63–66.
20. Conant GC, Wagner A (2004) Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans. Proc Biol Sci 271: 89–96.
21. Vilella A, Severin J, Ureta-Vidal A, Heng L, Durbin R, et al. (2009) Ensembl Compara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19: 327.
22. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, et al. (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res 39: D945–D950.

Chapter 2

Graphical Notation and Terminology

This chapter gives the general terminology and concepts on graphical models and causality that will be used throughout this dissertation.

2.1 Directed Acyclic Graphs

2.1.1 Graphical model terminology

A graphical model $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a representation of the relationships between random variables represented by the set of vertices $\mathbf{V} = \{X_1, X_2, ..., X_q\}$. Each edge in $\mathbf{E}$ denotes a relation between a pair of variables. If all the pairs of variables are connected by an edge, the graph is said to be *complete*. The edges can either be *directed* ($\rightarrow$) or *undirected* ($-$). When all edges are directed, the graph is also said *directed*. Graphical models that include both directed and undirected edges are named *partially directed*. When replacing all directed edges of a (partially) directed graph, one obtains a fully undirected network called *skeleton*.

Each vertex X_i can be assigned an *adjacency set*, $adj_{\mathcal{G}}(X_i)$, corresponding to the set of vertices that are connected by an edge to X_i in $\mathcal{G}$. If there exists a directed relation between X_i and X_j, such that $X_j \rightarrow X_i$, then X_j is called a *parent* of X_i. A chain of connected vertices constitutes a *path*. If all edges included in a path are oriented in the same direction (*e.g.* $\leftarrow\leftarrow$), the path is said to be *directed*. Whenever the path is directed from X_i to X_j and there exists a directed edge $X_j \rightarrow X_i$, then the path forms a *directed cycle*. If a (partially) directed graphical

model does not contain any directed cycle, it is called a (Partially) Directed Acyclic Graph ($(P)DAG$).

In (P)DAG, some triples $\langle X_i, X_k, X_j \rangle$ can be *unshielded*, meaning that only two pairs of nodes are adjacent, *e.g.* $X_i - X_k$ and $X_k - X_j$ while $X_i \neq X_j$. If an unshielded triple, $\langle X_i, X_k, X_j \rangle_{X_i \neq X_j}$, is oriented as $X_i \rightarrow X_k \leftarrow X_j$, it is said to be a *v-structure*. All the other possible orientations correspond to *non-v-structures* (Fig. 2.1).

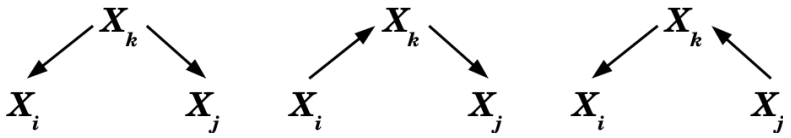


FIGURE 2.1 **Possible non-v-structures for an unshielded triples.**
The three possible non v-structures of the unshielded triples, $\langle X_i, X_k, X_j \rangle_{X_i \neq X_j}$, entail the same conditional independence, *i.e.* $X_i \perp\!\!\!\perp X_j | X_k$.

As detailed in the following sections, the three possible *non-v-structures* of an unshielded triple $\langle X_i, X_k, X_j \rangle_{X_i \neq X_j}$ entail the same conditional independence information (*i.e.* $X_i \perp\!\!\!\perp X_j | X_k$), and thus cannot be differentiated from one another, based solely on observational data.

2.1.2 Interpretation of a DAG

The set of conditional independence relationships among the variables $\mathbf{V}$, entailed by a DAG $\mathcal{G}$, can be deduced by applying a graphical criterion called the *d-separation* [Pearl, 2000] (see section 2.1.2.1). If two sets of vertices $(\mathbf{X_i}, \mathbf{X_j})$ are not adjacent in a DAG $\mathcal{G}$, they are *d-separated* by a subset $\mathbf{S}$ of the remaining variables in $\mathcal{G}$. It follows that $\mathbf{X_i}$ and $\mathbf{X_j}$ are independent conditional on $\mathbf{S}$, *i.e.*, $\mathbf{X_i} \perp\!\!\!\perp \mathbf{X_j} | \mathbf{S}$, in any distribution $\mathcal{P}$ over $\mathbf{V}$ that *factorizes*[1] according to $\mathcal{G}$. In fact, when assuming that the DAG $\mathcal{G}$ has the *local Markov* property[2], we assume that the conditional independence relationships obtained *via* the *d-separation* criterion *at least* holds in the distribution $\mathcal{P}$ over $\mathbf{V}$. Conversely, if the conditional independencies in $\mathcal{P}$ correspond *exactly*[3] to the ones that can be read off from $\mathcal{G}$ *via* the *d-separation* criterion, $\mathcal{P}$ is said to be *faithful* to $\mathcal{G}$ or *stable*.

[1] $\mathcal{P}$ *factorizes* according to a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ if the joint density distribution of $\mathbf{V}$ can be written as $p(X_1, X_2, ..., X_q) = \prod_{i=1}^{q} p(X_i | pa(\mathcal{G}, X_i))$, where $pa(\mathcal{G}, X_i)$ are the parents of X_i according to $\mathcal{G}$.
[2] Each vertex in the DAG $\mathcal{G}$ is independent of its non-descendants conditional on its parents.
[3] *i.e.* no additional independences can be found from the population

2.1.2.1 The d-separation criterion

This *d-separation* criterion is tied to the notion of *blocking a path* in a causal graphical model, *i.e.* stopping the flow of information between two related variables. In particular, if three nodes (X_i, X_k, X_j) are related through a *non-v-structure* (either as a *chain* $X_i \rightarrow X_k \rightarrow X_j$ or a *fork* $X_i \leftarrow X_k \rightarrow X_j$), the two extreme nodes are *marginally dependent*. However, when one knows the value of X_k, learning something about X_i has no effect on the probability of X_j. Thus, once the value of X_k is known, X_i and X_j become *independent*. In other words, X_k *blocks* the path between X_i and X_j.

On the other hand, if (X_i, X_k, X_j) are related through a *v-structure* (a *collider* $X_i \rightarrow X_k \leftarrow X_j$), the two extreme variables are, *a priori*, *marginally independent*. Yet, X_i and X_j become *dependent* conditioned on the value of X_k. As an example, we could consider that X_i represents the possibility of an *earthquake*, X_j denotes the possibility of *being burgled* and X_k stands for the possible activation of an *house alarm* (due to an earthquake or a burglary). *A priori*, no relation exists between the occurrence of an earthquake (X_i) and the fact of being robbed (X_j) (assuming that the integrity of the house is preserved). However, let's suppose that the activation of the house alarm is acknowledged (*i.e.*, the value of X_k is known). If the occurrence of an earthquake is then ascertained, one can draw the conclusion that the house is (likely) safe and, rather than a burglar, the earthquake has induced the activation of the alarm. This effect is also known as the *explaining away effect*[4].

A definition of the *d-separation* can be taken from [Pearl, 2000, Definition 1.2.3],

Definition 2.1.1. d-Separation
A *path* p is said to be *d-separated* (or *blocked*) by a set of nodes $\mathbf{Z}$ if and only if

1. p contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that the middle node m is in $\mathbf{Z}$, or
2. p contains an inverted fork (or *collider*) $i \rightarrow m \leftarrow j$ such that the middle node m is not in $\mathbf{Z}$ and such that no descendant of m is in $\mathbf{Z}$.

A set $\mathbf{Z}$ is said to *d-separate* $\mathbf{X}$ from $\mathbf{Y}$ if and only if $\mathbf{Z}$ blocks every path from a node in $\mathbf{X}$ to a node in $\mathbf{Y}$.

[4] This example has been taken from [Naïm et al., 2011, Chapter 1]

2.1.2.2 Markov equivalence

Several causal DAGs can entail the same conditional independence relationships, as introduced in section 2.1.1 for the three possible *non-v-structures* of an unshielded triple (Fig. 2.1). These DAGs are said to be *Markov equivalent* and the class they populate is called a *Markov equivalence class*. More specifically, two DAGs are equivalent if and only if they have the same *skeleton* and the same *v-structures* [Verma and Pearl, 1991].

A Markov equivalence class can be *uniquely* represented by a *Completed Partially Directed Acyclic Graph* (CPDAG) [Andersson et al., 1997, Chickering, 2002] that has the following properties:

1. every edge of a CPDAG exists in each DAG of the Markov equivalence class
2. every undirected edge of a CPDAG is oriented as $\rightarrow$ in at least one DAG and as $\leftarrow$ in at least one other DAG of the Markov equivalence class

If one assumes *faithfulness* and *causal sufficiency* (*i.e.*, no unmeasured common causes and no unmeasured selection variables, see section 2.2), a CPDAG $\mathcal{C}$ can be learned from the conditional independences between the observed random variables and used to represent the equivalence class of DAGs of the underlying causal graphical model.

2.2 Ancestral Graphs

As introduced in section 2.1, a CPDAG can be learned from the conditional independence information for systems satisfying the *causal sufficiency* assumption and the *faithfulness* assumption. Yet, in practice, the *causal sufficiency* assumption is rarely satisfied and a CPDAG may not properly represent the underlying causal structure among the observed variables, unless possible hidden common causes, or *latent variables*, are explicitly considered. However, explicitly introducing all possible latent variables, without any constraints on the network topology or on the number of hidden common causes, creates an infinite search space of DAGs [Zhang, 2008a]. To circumvent this intractable problem, the class of *ancestral graph models* has been introduced [Richardson and Spirtes, 2002]. As detailed below, this class of graphical models preserves the *ancestral* relationships encoded in the *true* underlying DAG, despite the presence of latent or selection variables.

The section 2.2.1 briefly describes common complications due to unobserved variables, while the section 2.2.2 gives the graphical terminology of ancestral graphs.

2.2.1 Complications from latent and selection variables

Two impediments can lead to the inference of spurious relationships when reconstructing a causal structure: the hidden (or latent) variables and the selection variables. The first ones correspond to unmeasured variables that contribute to observed associations. The second ones correspond to unobserved variables that characterize the subpopulation related to the samples drawn from the population of interest. As exemplified in section 2.2.1.1 by the variable Sel, the selection bias corresponds to a *by design* issue. Thus, any observed independence or association should always be regarded as *conditional on the selection variables*[5].

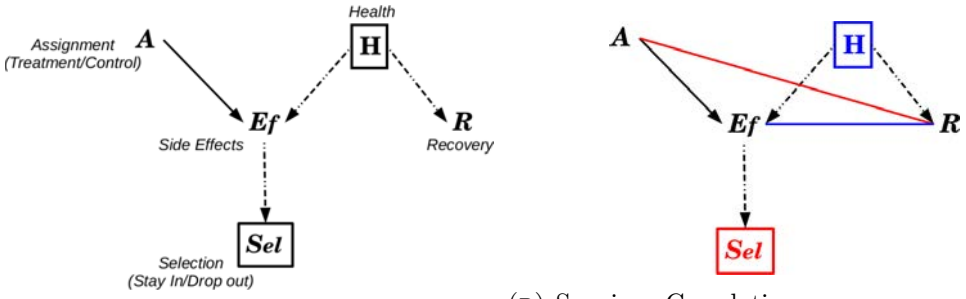
2.2.1.1 Spurious correlations

The following example, cited in [Richardson, 1999] and attributed to Chris Meek, describes how the presence of latent confounders and selection variables can induce 'spurious correlations' among the observed variables.

> *Randomized trial of an ineffective drug with unpleasant side-effects*
> 'The graph 2.2a represents a randomized trial of an ineffective drug with unpleasant side-effects. Patients are randomly assigned a treatment or control group (A). Those in the treatment group suffer from unpleasant side-effects (Ef), the severity of which is influenced by the patient's general level of health (H), with sicker patients suffering worse side-effects. Those patients who suffer sufficiently severe side-effects are likely to drop out of the study. The selection variable (Sel) records whether or not a patient remains in the study, thus for all those remaining in the study $Sel = StayIn$. Since unhealthy patients who are taking the drug are more likely to drop out, those patients in the treatment group who remain in the study tend to be healthier than those in the control group. Finally health status (H) influences how rapidly the patient recovers (R)'.

In Fig. 2.2, H is a hidden common cause responsible for the spurious correlation between Ef and R (blue edge). The selection variable Sel implies a correlation

[5] Although the effect of a selection variable results in a genuine probabilistic association, the induced association will be qualified as 'spurious' in the following, as for the association resulting from a hidden variable.



(A) Causal mechanism (B) Spurious Correlations

FIGURE 2.2 **Spurious correlations due to latent confounders and selection variables.** $\{A, Ef, R\}$ are the observed variables, H is a latent variable and Sel is a selection variable. In Fig. 2.2b, the red edge indicates the 'spurious correlation' due to the selection variable Sel, while the blue edge indicates the 'spurious correlation' induced by the hidden common cause H.

between A and R as the relationship between these two observed variables is in fact considered *condition* on Sel ($Sel = StayIn$).

2.2.1.2 Spurious causal effects

A related issue is the discovery of 'spurious causal effects'. In the DAG of Fig. 2.3, $X_i \perp\!\!\!\perp X_j$ is the only independence among the observed variables, leading to believe that the underlying causal structure over the observed variables $\mathbf{O}$ is the v-structure $X_i \rightarrow X_k \leftarrow X_j$. This CPDAG suggests that X_i and X_j are causes of X_k, while in fact no directed path exists between them. As detailed in section 2.2.2, the class of ancestral graphs would give a better representation of the relationships among the observed variables with the graphical model $X_i \longleftrightarrow X_k \longleftrightarrow X_j$, which is a *maximal ancestral graph* (MAG) that *uniquely* represents the underlying DAG of the Fig. 2.3 over the observed variables while taking into account the hidden variables. In this representation, the '>' mark should be read as 'not an ancestor of'. Thus, the '>' at X_i or X_j indicate that X_i or X_j are not the cause of X_k or of a selection variable. Similarly, the '>' around X_k indicates that X_k is not the cause of X_i, X_j or of a selection variable.

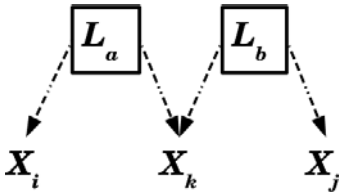


FIGURE 2.3 **An example of DAG with latent variables.** $\{X_i, X_k, X_j\}$ are observed variables, while $\{L_a, L_b\}$ are latent variables.

2.2.2 Graphical model terminology

The following sections give further details on a particular subclass of the class of *mixed graphs*, namely the class of *ancestral graphs*, introduced in section 2.2.1.2. A mixed graph has possibly two kinds of endpoints, *arrowhead* ($>$) and tail ($-$), and can thus contain three types of edge, *directed* ($\leftarrow$), *bi-directed* ($\longleftrightarrow$) and *undirected* ($—$). The symbol $*$ is used to represent any of the endpoint marks. In a mixed graph, two vertices X_i and X_j linked by any kind of edge are *adjacent*; X_i is a parent of X_j if $X_j \leftarrow X_i$; X_i is a spouse of X_j if $X_j \longleftrightarrow X_i$ and X_i is a neighbour of X_j if $X_j - X_i$. A chain of adjacent vertices is a *path*. A path from a vertex X_0 to the vertex X_p is said *directed* if $\forall\ 0 \leq i < p$, X_i is a parent of X_{i+1}. If X_p is a parent of X_0, the path forms a *directed cycle*. If the directed cycle contains a bi-directed edge, it is an *almost directed cycle*. X_i is an *ancestor* of X_j, $an_{\mathcal{G}}(X_j)$, if there is a directed path from X_i to X_j. X_j is then called a descendant of X_i, $de_{\mathcal{G}}(X_i)$. As will be seen in section 2.2.3.1, if we assume that the data have been generated by a DAG, and if we assume further the presence of latent and selection variables, then one can *uniquely* represent the conditional independencies and the causal relations entailed by the underlying DAG with a maximal ancestral graph (MAG) over the observed variables [Zhang, 2008a, page 1874].

The definition 2.2.1 gives the conditions for a mixed graph to be an ancestral graph.

Definition 2.2.1. Ancestral graph
A mixed graph $\mathcal{G}$ is an *ancestral graph* if:

1. $\mathcal{G}$ has no directed cycle.
2. $\mathcal{G}$ has no almost directed cycle.
3. For any undirected edge $X_i - X_j$, X_i and X_j have no parents or spouses.

The first condition can directly be related to the obligate absence of directed cycle in a DAG. The second condition, coupled with the first one, emphasizes the broader interpretation of the arrowhead mark in an ancestral graph, *i.e.* the arrowhead represents a *non-cause*. For instance, in the ancestral graph $X_i \rightarrow X_j$, the arrowhead at X_j indicates that X_j 'is not the cause' of X_i. By contrast, that tail at X_i gives a 'causal' information as it represents the fact that X_i 'is the cause' of X_j or of some selection variable. Hence, the tail is less informative than the arrowhead in the presence of selection bias. The third condition allows

a proper representation of the selection effect while simplifying the fitting of ancestral graphs [Richardson and Spirtes, 2002]. In a nutshell, an ancestral graph preserves, among the observed variables, the ancestral relationships encoded in the underlying DAG.

2.2.3 Interpretation of a MAG

As introduced earlier, when assuming the presence of hidden variables, a *maximal ancestral graph* (MAG) can uniquely entail a set of conditional independencies over the observed variables that can be deduced by applying the *m-separation* graphical criterion which generalizes the *d-separation* criterion defined in section 2.1 for DAGs.

2.2.3.1 The m-separation criterion

The *m-separation* criterion can be defined as follows:

Definition 2.2.2. m-Separation
A *path* p in an ancestral graph is said to be *m-separated* (or *blocked*) by a set of nodes $\mathbf{Z}$ if and only if

1. p contains a subpath $\langle X_i, X_j, X_k \rangle$ such that the middle vertex X_j is a *non-collider* on this path and $X_j \in \mathbf{Z}$, or
2. p contains a *collider*[6] $X_i * \!\!\rightarrow X_j \leftarrow \!\!* X_k$ such that $X_j \notin \mathbf{Z}$ and no descendant of X_j is in $\mathbf{Z}$.

A set $\mathbf{Z}$ is said to *m-separate* $\mathbf{X}$ from $\mathbf{Y}$ if and only if $\mathbf{Z}$ blocks every path from a node in $\mathbf{X}$ to a node in $\mathbf{Y}$. If all pairs of vertices $(X_i, X_k) \in \mathbf{X} \times \mathbf{Y}$ are *m-separated* by $\mathbf{Z}$, the sets of vertices $\mathbf{X}$ and $\mathbf{Y}$ are said *m-separated* by $\mathbf{Z}$.

Similarly as for DAGs, the *Markov property* states that if X_i and X_j are m-separated by X_k, then X_i and X_j are probabilistically independent conditional on X_k, and thus, $X_i \not\perp X_i$ in $\mathcal{G}$. However, by contrast with the DAGs, the reverse is not always true for an ancestral graph. In fact, unless an ancestral graph has the

[6] A vertex X_j is called a *collider* if the two endpoint marks at this vertex are arrowheads, otherwise it is called a *non-collider*. In the mix graph $X_i \longleftrightarrow X_j \longleftrightarrow X_k$, X_j is a *collider*. However, in the mix graph $X_i \longleftrightarrow X_j \rightarrow X_k$, X_j is a *non-collider*.

maximality property, it misses the so-called *pairwise Markov* property[7], necessary to infer the skeleton.

Definition 2.2.3. Maximality
An ancestral graph is said to be *maximal* if for any two non-adjacent vertices, there is a set of vertices that m-separates them.

It can be shown that given any DAG $\mathcal{G}$ over $\mathbf{V} = \mathbf{O} \cup \mathbf{L} \cup \mathbf{S}$, where $\mathbf{O}$ is the set of observed variables, $\mathbf{L}$ is the set of latent variables, $\mathbf{S}$ is the set of selection variables and $\mathbf{O}, \mathbf{L}, \mathbf{S}$ are disjoint sets, there exists a unique maximal ancestral graph (MAG) $\mathcal{M}_{\mathcal{G}}$ over $\mathbf{O}$ alone that *probabilistically* represents $\mathcal{G}$ [Richardson and Spirtes, 2002], such that:

$$\forall \text{ disjoint sets of variables } \mathbf{X}, \mathbf{Y}, \mathbf{Z},$$
$$\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{O}, \quad \mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z} \cup \mathbf{S} \text{ in } \mathcal{G} \iff \mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z} \text{ in } \mathcal{M}_{\mathcal{G}}.$$

Such MAG $\mathcal{M}_{\mathcal{G}}$ encodes the following causal information:

- $\mathbf{X} \to \mathbf{Y}$, *i.e.* $\mathbf{X} \in an_{\mathcal{M}_{\mathcal{G}}}(\mathbf{Y} \cup \mathbf{S})$ and $\mathbf{Y} \notin an_{\mathcal{M}_{\mathcal{G}}}(\mathbf{X} \cup \mathbf{S})$

  $\mathbf{X}$ *is a cause* of $\mathbf{Y}$ or of some selection variable (tail at $\mathbf{X}$), but $\mathbf{Y}$ *is not a cause* of $\mathbf{X}$ or of any selection variable (arrowhead at $\mathbf{Y}$)

- $\mathbf{X} \leftrightarrow \mathbf{Y}$, *i.e.* $\mathbf{X} \notin an_{\mathcal{M}_{\mathcal{G}}}(\mathbf{Y} \cup \mathbf{S})$ and $\mathbf{Y} \notin an_{\mathcal{M}_{\mathcal{G}}}(\mathbf{X} \cup \mathbf{S})$

  $\mathbf{X}$ *is not a cause* of $\mathbf{Y}$ or of any selection variable (arrowhead at $\mathbf{X}$), and $\mathbf{Y}$ *is not a cause* of $\mathbf{X}$ or of any selection variable (arrowhead at $\mathbf{Y}$), *i.e.* there is a common cause to $\mathbf{X}$ and $\mathbf{Y}$

- $\mathbf{X} - \mathbf{Y}$, *i.e.* $\mathbf{X} \in an_{\mathcal{M}_{\mathcal{G}}}(\mathbf{Y} \cup \mathbf{S})$ and $\mathbf{Y} \in an_{\mathcal{M}_{\mathcal{G}}}(\mathbf{X} \cup \mathbf{S})$

  $\mathbf{X}$ *is a cause* of $\mathbf{Y}$ or of some selection variable (tail at $\mathbf{X}$), and $\mathbf{Y}$ *is a cause of* $\mathbf{X}$ or of some selection variable (tail at $\mathbf{Y}$), *i.e.* because of the acyclicity assumption, $\mathbf{X}$ is a cause of some selection variable, and $\mathbf{Y}$ is a cause of some selection variable

2.2.3.2 Markov equivalence

As introduced in the section 2.1.2.2, two DAGs that entail the same d-separation structure, although they may encode different causal information, are Markov

[7] Every missing edge corresponds to a conditional independence relation

equivalent (*e.g.* Fig. 2.1). Similarly, one can define the Markov equivalence of two MAGs as follow:

Definition 2.2.4. Makov equivalence of two MAGs
Two MAGs $\mathcal{M}_{\mathcal{G}_1}, \mathcal{M}_{\mathcal{G}_2}$ are Markov equivalent if $\forall$ disjoint sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{O}$, $\mathbf{X}$ and $\mathbf{Y}$ are m-separated by $\mathbf{Z}$ in $\mathcal{M}_{\mathcal{G}_1} \iff \mathbf{X}$ and $\mathbf{Y}$ are m-separated by $\mathbf{Z}$ in $\mathcal{M}_{\mathcal{G}_2}$.

Hence, all the members of the Markov equivalence class of a given MAG have the same adjacencies and the same *invariant* endpoints (*arrowhead* ($>$) or tail ($-$)). Indeed, within the Markov equivalent class of MAGs, one or both endpoints of some edges may be different depending on the MAG member. The mark '$\circ$' indicates such variant endpoints.

The invariant endpoints and the edges learned from the conditional independencies can be represented by a *Partial Ancestral Graph* (PAG), having possibly six kinds of edges: $-$, $\rightarrow$, $\longleftrightarrow$, $-\circ$, $\circ\!-\!\circ$ and $\circ\!\rightarrow$. The bi-directed edges highlight the existence of hidden variables, and the undirected edges are the result of the selection variables.

The following gives the definition of a PAG:

Definition 2.2.5. Partial Ancestral Graph (PAG)
Let $\mathcal{G}$ be a DAG over $\mathbf{V} = \mathbf{O} \cup \mathbf{L} \cup \mathbf{S}$ –with $\mathbf{O}$ the set of observed variables, $\mathbf{L}$ the set of latent variables, $\mathbf{S}$ the set of selection variables and $\mathbf{O}, \mathbf{L}, \mathbf{S}$ disjoint sets–. Let $\mathcal{M}$ be a mixed graph over $\mathbf{O}$. $\mathcal{M}$ is said to be a PAG that represents $\mathcal{G}$ if and only if, for any distribution $\mathcal{P}$ over $\mathbf{V}$ that is faithful to $\mathcal{G}$, the following four conditions hold:

(i) $X_i \not\!-\!\!\!\!\circ\, X_j$ in $\mathcal{M} \Rightarrow \exists\, \mathbf{Y} \subseteq \mathbf{O} \setminus \{X_i, X_j\}$ such that $X_i \perp\!\!\!\perp X_j | \mathbf{Y} \cup \mathcal{S}$ in $\mathcal{P}$;

(ii) $X_i - X_j$ in $\mathcal{M} \Rightarrow \forall\, \mathbf{Y} \subseteq \mathbf{O} \setminus \{X_i, X_j\}, X_i \not\perp\!\!\!\perp X_j | \mathbf{Y} \cup \mathcal{S}$ in $\mathcal{P}$;

(iii) $X_i \circ\!\!\rightarrow X_j$ in $\mathcal{M} \Rightarrow$ then $\mathbf{X_j} \notin an_{\mathcal{M}}(\mathbf{X_i} \cup \mathbf{S})$

(iv) $X_i \circ\!\!- X_j$ in $\mathcal{M} \Rightarrow$ then $\mathbf{X_j} \in an_{\mathcal{M}}(\mathbf{X_i} \cup \mathbf{S})$

As an example, the graphical model $X_i \circ\!\!\rightarrow X_k \leftarrow\!\!\circ X_j$ is the PAG over the observed variables $\mathbf{O}$ that represents the Markov equivalence class of MAGs of the underlying causal graphical model of Fig. 2.3. In this representation, the '$>$' around X_k indicates that X_k is not the cause of X_i, X_j or a selection variable, and the '$\circ$' mark at X_i or X_j indicates uncertainty about the possibility of X_i and X_j to be a cause of X_k.

2.2.3.3 Finding adjacencies in a PAG

As introduced in the section 2.2.1.1, the hidden variables prevent from inferring conditional independencies among the observed variables. The example of Fig. 2.4, taken from [Spirtes and Glymour, 1991], illustrates a spurious causal dependence due to the presence of latent variables.

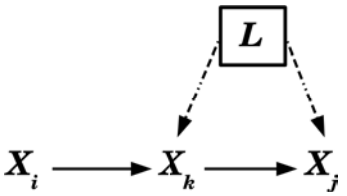


FIGURE 2.4 **Example of a DAG $\mathcal{G}$ over $\mathbf{V} = \mathbf{O} \cup \mathbf{L} \cup \mathbf{S}$.** $\{X_i, X_k, X_j\} \subseteq \mathbf{O}$ where $\mathbf{O}$ is the set of observed variables and $\{L\} \subseteq \mathbf{L}$ where $\mathbf{L}$ is the set of latent variables.

The latent variable L prevents from inferring the independence between X_i and X_j conditional on X_k over the observed variables $\mathbf{O}$, and hence from the removal of the edge between X_i and X_j, although X_i has no direct influence on X_j in the DAG $\mathcal{G}$ over $\mathbf{V} = \mathbf{O} \cup \mathbf{L} \cup \mathbf{S}$. Yet, following the definition of a PAG $\mathcal{M}$ (see definition 2.2.5), one can still decide for the adjacency between two observed variables when accounting for hidden variables, as:

$$\left.\begin{array}{r} X_i \notin an_{\mathcal{M}}(X_j \cup \mathbf{S}) \\ \text{and,} \\ X_i \perp\!\!\!\perp X_j | \mathbf{Y} \cup \mathbf{S} \end{array}\right\} \Rightarrow X_i \perp\!\!\!\perp X_j | \mathbf{Y}' \cup \mathbf{S}$$

with,

- $\mathbf{Y} \subseteq \mathbf{O} \setminus \{X_i, X_j\}$
- $\mathbf{Y}' \subset$ D-SEP$\{X_i, X_j\}$ or $\subset$ D-SEP$\{X_j, X_i\}$

Hence, to decide whether X_i and X_j are adjacent in the PAG $\mathcal{M}$, it is sufficient to find a separation set $\mathbf{Y}'$ from a given D-SEP of $\{X_i, X_j\}$ rather than trying all possible subsets in $\mathbf{Y} \subseteq \mathbf{O} \setminus \{X_i, X_j\}$. However, the D-SEP sets — which definition is closely related to the notion of *inducing path* — cannot be deduced from the observed conditional independencies [Spirtes and Glymour, 1991, Spirtes et al., 1999]. Hence, a superset containing the D-SEP, called Possible-D-SEP, has been proposed by [Spirtes et al., 2000].

A Possible-D-SEP set is defined over the observed variables $\mathbf{O}$ as,

Definition 2.2.6. Possible-D-SEP
Let the distribution $\mathcal{P}$ over $\mathbf{V} = \mathbf{O} \cup \mathbf{L} \cup \mathbf{S}$ be faithful to a DAG $\mathcal{G}$. Let $\mathcal{C}$ be a PAG that represents $\mathcal{G}$. A vertex X_k belongs to *Possible-D-SEP*(X_i, X_j), $pds_{\mathcal{C}}(X_i, X_j)$, if and only if there is a path π between X_i and X_j in $\mathcal{C}$ such that for every subpath $\langle X_m, X_l, X_h \rangle$ of π, X_l is a collider on the subpath in $\mathcal{C}$ or $\langle X_m, X_l, X_h \rangle$ is a triangle in $\mathcal{C}$.

As detailed in the section 3.2.2.1, these Possible-D-SEP set of variables can be used to learn the representative of the Markov equivalent class of DAGs of the underlying causal graphical model when allowing for the presence of latent and selection variables.

Given observational data, the problem of learning a DAG or a MAG (or, a representative of the corresponding Markov equivalence class, *i.e.* a CPDAG or a PAG, resp.) can be solved by two main types of methods relying either on a *contraint-based* approach (see Chapter 3), or on a *search-and-score* process (see Chapter 4). Hybrid approaches, based on a combination of these two types of methods are also discussed in the following chapters, as well as our new hybrid causal discovery method, namely 3off2, detailed in Chapter 6.

Chapter 3

Constraint-based Methods

Constraint-based approaches, such as the PC ([Spirtes and Glymour, 1991, Spirtes et al., 2000]) and IC ([Pearl and Verma, 1991]) algorithms, learn causal graphical models from observational data by searching for conditional independencies among variables. When assuming that the data have been generated by a DAG model, these algorithms return a Complete Partially Directed Acyclic Graph (CPDAG) that represents the Markov equivalent class of the underlying causal structure under the *faithfulness*[1] and *causal sufficiency*[2] assumptions [Pearl and Verma, 1991, Spirtes et al., 2000]. To ascertain conditional independences between variables, these methods usually rely on statistical tests that depend on a significance level α. These approaches based on the identification of structural constraints typically run in polynomial time on sparse underlying graphs and are computationally feasible for large graphical models, up to thousands of variables [Kalisch and Bühlmann, 2007]. This chapter gives an overview of the constraint-based approach principles.

> The novel approach and corresponding 3off2 algorithm, that will be detailed in Chapter 6, is directly inspired from the constraint-based methods. However, it relies on a quantitative information theoretic framework that allows for a robust identification of conditional independencies.

[1] The independence relations that are read off the DAG $\mathcal{G}$ using the *d-separation* graphical criterion hold in the probability distribution over $\mathbf{O} = \{X_1, X_2, ..., X_q\}$ and reciprocally (see also subsection 2.1.2.1)

[2] The set of *measured* or *observed* variables $\mathbf{O}$ contains all the common causes of pairs of variables, *i.e.* there are no hidden confounders and no hidden selection variables

3.1 The PC algorithm

To reconstruct the underlying causal graphical model $\mathcal{G} = (\mathbf{O}, \mathbf{E})$, where $\mathbf{O}$ is the set of observed variables and $\mathbf{E}$ the set of edges, constraint-based methods, and in particular the widely used PC ([Spirtes and Glymour, 1991, Spirtes et al., 2000]) algorithm, typically proceed in three steps:

1) learning unnecessary edges and separation sets to obtain an undirected skeleton (Algorithm 1, lines $[7-17]$)

2) orienting unshielded triples as v-structures if their middle node is not in the separation set (Algorithm 1, rule R_0, lines $[19-21]$)

3) propagating as many orientations as possible following propagation rules (R_{1-3}), which prevent the orientation of additional v-structures (R_3) and directed cycles (R_{2-3}) (Algorithm 1, lines $[23-27]$)

The following sections give further details on each of these steps.

3.1.1 Learning the skeleton and sepsets

In the first step of the PC algorithm (Algorithm 1, (lines $[7-17]$)), the complete undirected graphical model taken as input is progressively thinned out based on the discovery of conditional independence relationships.

The efficiency of the PC algorithm lies in the organization of the conditional independence queries, $Indep(X; Y|\{U_i\})$ (Algorithm 1, (line [10])). Instead of checking all the conditional independencies, as the SGS algorithm does [Spirtes et al., 2000], the PC algorithm is searching for the separation sets among the remaining adjacent vertices during the skeleton thinning out process, which can lead to a polynomial time execution in graphs of finite degree. First, starting with an empty separation set $\ell = 0$, all pairs of variables, (X, Y), are tested for *marginal independence*. Whenever such independence is found, $X \perp\!\!\!\perp Y$, the corresponding edge is removed, $X \neq Y$, and the empty set is stored as separation set, $\text{Sep}_{XY} = \text{Sep}_{YX} = \emptyset$. The value of ℓ is then incremented, and each remaining edge having vertices with an adjacency set of size at least ℓ is tested for conditional independence based on subset $\{U_i\} \subseteq adj_{\mathcal{G}}(X) \setminus \{Y\}$ (or $\{U_i\} \subseteq adj_{\mathcal{G}}(Y) \setminus \{X\}$). The PC algorithm iterates until all adjacency sets are smaller than ℓ. This first step returns the skeleton of the underlying graphical model and a separation set (possibly empty) for each missing edge.

As will be shown later, a major shortcoming of the PC algorithm is that it is not robust to sampling noise in finite dataset, which leads to the accumulation of errors during the edges pruning procedure.

By contrast, as will be shown in Chapter 6, the 3off2 learning method uncovers the conditional independencies among the observed variables following a robust iterative approach in order to build the skeleton of the underlying graphical model (section 6.3). In a nutshell, the 3off2 process iteratively builds the separation sets guided by a score based on 2-point and 3-point information statistics (Eq. 6.32) until no variable belonging to the separation sets can be found.

3.1.2 Orientating the skeleton

Once the skeleton and the separation sets are known, the v-structures can be identified (Algorithm 1, lines $[19-21]$). As introduced in the subsection 2.1.2.1, if the unshielded triple $\langle X_i, X_k, X_j \rangle_{X_i \not\sim X_j}$ is a v-structure, knowing the value of X_k creates a dependence between X_i and X_j, while if $\langle X_i, X_k, X_j \rangle_{X_i \not\sim X_j}$ is a non-v-structure, knowing the value of X_k induces a conditional independence between X_i and X_j. Thus, if a separation set between X_i and X_j has been found, and if X_k does not belong to it, then under the faithfulness and the causal sufficiency assumptions, one can conclude that $\langle X_i, X_k, X_j \rangle_{X_i \not\sim X_j}$ is a v-structure and orient the two edges accordingly.

By contrast, as will be shown in section section 5.1.3, the 3off2 approach relies on the sign and magnitude of the 3-point information in order to decide whether a unshielded triple could be oriented as a collider (section 6.3.2.2).

3.1.3 Propagating the orientations

The last step of the PC algorithm consists in *propagating* the orientations to the remaining undirected edges following propagation *rules* (Algorithm 1, lines $[23-27]$) that guarantee that no directed cycles or new v-structures will be created. It has been shown that repeated application of the three rules R_{1-3} is sufficient to recover the CPDAG of the graphical model [Meek, 1995].

By contrast, in the 3off2 algorithm, the propagation and the orientation steps are simultaneously performed, following the rank of each unshielded triple (section 6.3.2.2).

Algorithm 1: The PC algorithm

In: observational data of variables $\mathbf{O}$; an ordering $order(\mathbf{O})$ on the variables; a confidence level α

Out: CPDAG $\mathcal{C}$

0. Initiation

Start with a complete undirected graph $\mathcal{G}$

Let $\ell = 0$

1. Iteration

repeat

  while $\exists$ XY *link with* $|adj_{\mathcal{G}}(X)\backslash\{Y\}| \geqslant \ell$ **do**

    while $XY \subset adj_{\mathcal{G}}$ and $\exists\{U_i\} \subseteq adj_{\mathcal{G}}(X)\backslash\{Y\}$, *not yet considered with* $|\{U_i\}| = \ell$ **do**

      if $\text{Indep}(X;Y|\{U_i\})$ *at conf. level* α **then**

        XY **link is** non-essential and **removed**

        separation set of XY: $\text{Sep}_{XY} = \{U_i\}$

      end

    end

  end

  Set $\ell = \ell + 1$

until $\forall X \in \mathcal{G}$, $|adj_{\mathcal{G}}(X)| \leqslant \ell$;

2. Orientation

forall the *unshielded triples* **do**

  R_0: $\{X - Z - Y \;\&\; X \neq Y \;\&\; Z \notin \text{Sep}_{XY}\} \;\Rightarrow \{X \to Z \leftarrow Y\}$

end

3. Propagation

repeat

  R_1: $\{X \to Z - Y \;\&\; X \neq Y\} \qquad \Rightarrow \{Z \to Y\}$

  R_2: $\{X \to Y \to Z \;\&\; X - Z\} \qquad \Rightarrow \{X \to Z\}$

  R_3: $\{X - Y \to Z \;\&\; X - T \to Z \;\&\; Y \neq T\} \qquad \Rightarrow \{X \to Z\}$

until *no further orientation can be propagated*;

3.1.4 Statistical tests for conditional independence

The conditional independence tests performed during the pruning procedure of constraint-based methods, and in particular during the iteration step of the PC algorithm (Algorithm 1, line [10]), usually rely on χ^2 or G^2 independence tests. The decision of removing the edge between two variables X_i and X_j conditioned on a set of variables $\mathbf{X_k}$ is made through the confrontation of the following hypothesis:

$\mathbf{H_0}$ The variables X_i and X_j are independent conditioned on $\mathbf{X_k}$

$\mathbf{H_a}$ The variables X_i and X_j are not independent conditioned on $\mathbf{X_k}$

3.1.4.1 Chi-square conditional independence test

A χ^2 independence test can be used to accept or reject the null hypothesis, *i.e.* the independence between two discrete variables, X_i and X_j, conditioned on a set of discrete variables, $\mathbf{X_k}$ (where X_i has r_i levels, X_j has r_j levels, and there exists q_k combinations of levels for the set $\mathbf{X_k}$).

The test confronts the following models:

- the independence model: $p_i = P(X_i|\mathbf{X_k})P(X_j|\mathbf{X_k})$
- the observed model: $p_o = P(X_i, X_j|\mathbf{X_k})$

The model p_o is represented by N_{ijk}, the number of data points with $\{X_i = x_i$,$X_j = x_j$ and $\mathbf{X_k} = \mathbf{x_k}\}$ (with $\mathbf{x_k}$ the kth state of the set $\mathbf{X_k}$). The model p_i is represented by $\frac{N_{ik} \times Njk}{N_k}$, where N_{ik} corresponds to the number of data points with $\{X_i = x_i$ and $\mathbf{X_k} = \mathbf{x_k}\}$, N_{jk} corresponds to the number of data points with $\{X_j = x_j$ and $\mathbf{X_k} = \mathbf{x_k}\}$, and N_k corresponds to the number of data points with $\{\mathbf{X_k} = \mathbf{x_k}\}$.

The test statistics is a χ^2 random variable defined as follow:

$$\chi^2_{statistics} = \sum_{i,j,l}^{r_i} \sum_{j=1}^{r_j} \sum_{k=1}^{r_k} \frac{(N_{ijk} - \frac{N_{ik} \times Njk}{N_k})^2}{\frac{N_{ik} \times Njk}{N_k}} \tag{3.1}$$

Then, one can compute a *p-value* corresponding to the probability of observing a sample statistics as extreme as the test statistics, $\chi^2_{statistics}$. The independence

between X_i and X_j condition on X_k is rejected (*i.e.* the edge is not removed) for a significance level α if and only if:

$$P\Big(\chi^2_{statistics} < \chi^2\big(df\big)\Big) < \alpha \tag{3.2}$$

where $df = (r_i - 1)(r_j - 1)q_k$ is the the degree of freedom

3.1.4.2 G-square conditional independence test

Instead of the χ^2 statistics, Spirtes *et al.* proposed to use the G^2 likelihood ratio, that also follows a χ^2 with a degree of freedom $df = (r_i - 1)(r_j - 1)q_k$:

$$G^2 = 2\sum_{i=1}^{r_i}\sum_{j=1}^{r_j}\sum_{k=1}^{r_k} N_{ijk} ln\Big(\frac{N_{ijk}}{\frac{N_{ik}\times N_{jk}}{N_k}}\Big) \tag{3.3}$$

Similarly as for the χ^2 statistics, the G^2 likelihood ratio can the used to accept or reject the null hypothesis at a significance level α.

The significance level α may be thought as an adjustable parameter to improve the discovery of the underlying causal structure. A smaller α value tends to favour the independence model and thus to return a less connected graph. Conversely, a larger α tends to favour more complex causal graphical models, as can be seen in Fig. B.7 - B.12. Besides, as with large datasets the dependencies tend to have p-values approaching 0 while the p-value of independencies are uniformly distributed from 0 to 1, one should also lower the significance level when the sample size increases, as can be seen in Fig. A.1 - A.5.

In practice, the 3off2 structure learning approach relies instead on the minimization of a 2-point information term, and the simultaneous maximization of a 3-point information term in order to estimate the likelihood that the edge $X - Y$ should be removed (section 6.2.3, Eqs. 6.4 & 6.5). This strategy is shown to be more robust to sampling noise in finite datasets for benchmark networks (Chapter 7).

3.2 Variations on the PC algorithm

When an *exact* list of conditional independence relationships is given as input, the PC or the IC algorithms are guaranteed to learn the CPDAG of the underlying graphical model. However in practice, constraint-based methods should rely on statistical independence tests (subsection 3.1.4), to ascertain the conditional independencies from observational data.

As previously stated [Colombo and Maathuis, 2014, Dash and Druzdzel, 1999], the use of a statistical test on finite datasets makes the constraint-based methods not robust to sampling noise in finite datasets. Indeed, early errors in removing edges from the complete graph typically trigger the accumulation of compensatory errors later on in the prunning process and this cascading effect makes the constraint-based approaches sensitive to the adjustable significance level α required for the conditional independence tests. In addition, traditional constraint-based methods are not robust to the order in which the conditional independence tests are processed, possibly leading to a wide range of different results, in particular in high-dimensional settings. This prompted recent algorithmic improvements intending to achieve order-independence, as the *PC-stable* algorithm proposed by [Colombo and Maathuis, 2014], and tested on experimental observational data from yeast gene expression.

Besides, the PC algorithm, as well as other structure discovery approaches, requires specific assumptions to guarantee that the output would be the representative of the markov equivalent class of the underlying graphical model. Among these assumptions, the *causal sufficiency* plays an important role as it assumes the absence of *latent* and *selection* variables. Yet in practice, the system under observation typically contains unmeasured common causes and selection variables (section 2.2.1). The FCI ('Fast Causal Inference') algorithm [Spirtes et al., 1999, 2000], and the more efficient RFCI ('Really Fast Causal Inference') algorithm [Colombo et al., 2012], have been proposed to enable the discovery of the Partial Ancestral Graph (PAG) (section 2.2.3.2), representative of the Markov equivalence class of DAGs when accounting for hidden and selection variables. The following subsections detail and discuss the PC-stable, the FCI and the RFCI algorithms.

3.2.1 PC-stable: an order-independent constraint-based approach

The dependence on the order in which the variables are given impedes the skeleton learning and the edge orientation steps, in particular with high dimensional settings. [Colombo et al., 2012] proposed a modification of the PC algorithm that removes the order-dependence on the input variables while discovering the skeleton. [Ramsey et al., 2006] and [Colombo et al., 2012] also proposed complementary orientation procedures to reach order-independence during the orientation/propagation steps.

3.2.1.1 Order-independent skeleton learning

The order-dependence of the PC algorithm is related to the removal of the *non-essential* edge (Algorithm 1, lines [11]) *within* one level ℓ (Algorithm 1, lines [8 − 16]), where ℓ corresponds to the size of the separation set. Indeed, removing an edge at a given level ℓ possibly modifies the adjacency sets of the remaining edges not yet tested for conditional independence, and thus influences the outcome of the following statistical tests at this level. In practice, mistakes can occur when checking for conditional independence using a statistical test and lead to erroneous skeleton and separation sets. As the order in which the input variables are considered can imply different errors, a wide range of skeletons can be expected, in particular for high dimensional settings as shown in [Colombo and Maathuis, 2014].

[Colombo and Maathuis, 2014] proposed an order-independent version of the skeleton discovery step for the PC algorithm, called *PC-stable*, where the edge removal is not done *within* a specific level ℓ. Instead, the *PC-stable* algorithm stores all the possible separation sets found for a given level ℓ and removes the edges corresponding to the conditional independence only before reaching the next level (or eventually before ending the algorithm). This *postponed* edge removal *between* successive levels, rather than *within* one level, allows for an order-independent discovery of the skeleton, as the edges under statistical test for a level ℓ do not influence each others adjacency sets when removed.

As detailed in section 6.3.2, the 3off2 inference approach proceeds on a list of triples ordered by a rank (Eq. 6.32) that optimizes the likelihood that a variable Z is contributing to the relationship between two other variables, X and Y. As the 3off2 inference procedure relies on this rank, the skeleton produced is not

dependent on the order in which the input variables are processed.

3.2.1.2 Order-independent orientation steps

As exemplified in [Colombo et al., 2012], the information contained in the separation sets are also order-dependent, as some separating sets may hold by mistake at some level α. When used for the orientation of the v-structures, these spurious separation sets may lead to incorrect orientations depending on the order in which the given variables are processed.

[Ramsey et al., 2006] proposed a *conservative* rule to remove the order-dependence of this orientationstep. [Colombo and Maathuis, 2014] proposed a similar but less conservative approach. Both methods rely on the identification of *unambiguous* triples, and differ in the definition of the *unambiguity*. These approaches first collect, for each unshielded triples $\langle X_i, X_k, X_j \rangle_{X_i \neq X_j}$ in the reconstructed skeleton $\mathcal{C}$, all the separation sets $\{\mathbf{Y}\}$ among the variables belonging to $adj_{\mathcal{C}}(X_i)$ and $adj_{\mathcal{C}}(X_j)$, such that $X_i \perp\!\!\!\perp X_j | \mathbf{Y}$. For the conservative approach [Ramsey et al., 2006], $\langle X_i, X_k, X_j \rangle_{X_i \neq X_j}$ is said *unambiguous* if at least one separation set $\mathbf{Y}$ can be found, and X_k belongs to all or none of the separation sets $\{\mathbf{Y}\}$. When the triple is *unambiguous*, it is oriented as a v-structure if and only if X_k is in none of the separation sets $\{\mathbf{Y}\}$. By contrast, [Colombo and Maathuis, 2014] rely on a *majority* rule to decide for unambiguity. $\langle X_i, X_k, X_j \rangle_{X_i \neq X_j}$ is said unambiguous if at least one separation set $\mathbf{Y}$ can be found, and X_k is not in exactly 50% of the $\{\mathbf{Y}\}$. When the triple is unambiguous, it is oriented as a v-structure if and only if X_k is in less than half of the separation sets $\{\mathbf{Y}\}$. For both methods, only the unambiguous triples are considered in the orientation propagation step. The algorithms resulting from the association of the PC-stable step and the conservative or the majority rule orientation step are called respectively *CPC-stable* and *MPC-stable*.

Finally, even when applying the CPC-stable or the MPC-stable algorithms, some *orientation conflicts* may occur among the unambiguous v-structures, for instance when $X_1 \rightarrow X_2 \leftarrow X_3$ and $X_2 \rightarrow X_3 \leftarrow X_4$ have both been learned. To avoid order-dependence, one can resolve the orientation conflict by setting a *bi-directional* edge between X_2 and X_3[3]. The algorithms resulting from the association of the PC-stable step, the conservative or the majority rule orientation step, and the orientation conflict feature are called respectively LCPC-stable and LMPC-stable.

[3] As these bi-directed edges only represent conflicting information, [Colombo and Maathuis, 2014] stipulate that they cannot be interpreted causally (page 3756)

As detailed in the section 6.3.2.2, the 3off2 orientation/propagation step ranks the unshielded triples by the absolute value of their 3-point information. Then, depending on the sign of the 3-point information, the unshielded triples are successively oriented as a v-structure or allow for a propagation (Algorithm 4, step 2. and Algorithm 5). This procedure based, solely on statistically significant 3-point information, provides a robust orientation/propagation step, independent of the order in which the variables are processed.

3.2.2 Causal Inference with latent and selection variables

3.2.2.1 The FCI algorithm

As introduced in the section 2.2.1, unmeasured variables such as latent or selection variables, cause complications when learning a causal graphical model from the observed variables. The FCI ('Fast Causal Inference') algorithm [Spirtes et al., 1999, 2000] allows the discovery of the representative of the Markov equivalence class of DAGs, from observed conditional independence information, while allowing for the presence of latent and selection variables.

Algorithm 2 gives the details of the FCI algorithm. First, the discovery of the skeleton with the PC algorithm (Algorithm 2, step 1.) and the identification of the v-structures (Algorithm 2, step 2.) are performed. Then, complementary conditional independence tests, for which the conditioning set of variables is search only in the Possible-D-SEP (see section 2.2.3.3), are processed (Algorithm 2, step 3.). Before applying the final orientation rules (Algorithm 2, step 4.), the edges of the skeleton are all set to $\circ\!\!-\!\!\circ$. Finally, the propagation step follows the rules R_1 to R_{10} established in [Zhang, 2008a]. The returned graphical model $\mathcal{C}$, akin to a PAG [Zhang, 2008b], is said maximally informative.

The Fig. 3.1, taken from [Colombo et al., 2012] (section 3.3), illustrates the reconstruction of the PAG over $\mathbf{O}$ (Fig. 3.1, (c)) that represents the markov equivalence class of the DAG $\mathcal{G}$ over $\mathbf{V} = \mathbf{O} \cup \mathbf{L} \cup \mathbf{S}$ (Fig. 3.1, (a)). The causal independence relationships have been deduced from the observed variables, $\{X_i\}$ (Fig. 3.1).

As can be seen, the edge $X_1 \circ\!\!-\!\!\circ X_5$ has not been remove after the two first steps of the FCI algorithm as the vertex X_3, required for the conditional independence between X_1 and X_5, does not belong to $adj_{\mathcal{C}_1}(X_1)$ nor $adj_{\mathcal{C}_1}(X_5)$. However, in the following third step, the FCI algorithm searches more distant vertices as it relies on Possible-D-SEP to find the separation sets. Since $pds_{\mathcal{C}_2}(X_1, X_5) \setminus \{X_1, X_5\} =$

Algorithm 2: The FCI algorithm

In: observational data of variables $\mathbf{O}$; an ordering $order(\mathbf{O})$ on the variables; a confidence level α

Out: (a mixed graph interpretable as) a PAG $\mathcal{C}$ [Zhang, 2008b]

0. Initiation

Start with a complete undirected graph $\mathcal{C}$ with edges $\circ\!\!-\!\!\circ$

1. Initial skeleton, separation sets, out: $\mathcal{C}_1$

Do PC algorithm to find an initial skeleton $\mathcal{C}_1$ and the separation sets (Algorithm 1)

2. Initial orientations, out: $\mathcal{C}_2$

forall the *unshielded triples in* $\mathcal{C}_1$ **do**
  R_0: $\{X *\!\!-\!\!\circ Z \circ\!\!-\!\!* Y \;\&\; X *\!\!\not-\!\!* Y \;\&\; Z \notin \mathrm{Sep}_{XY}\} \;\Rightarrow \{X *\!\!\rightarrow Z \leftarrow\!\!* Y\}$
end

3. Final skeleton, separation sets, out: $\mathcal{C}_3$

for *all vertices* $X \in \mathcal{C}_2$ **do**
  Compute $pds_{\mathcal{C}_2}(X, \bullet)$
  for *all vertices* $Y \in adj_{\mathcal{C}_2}(X)$ **do**
    Let $\ell = 0$
    while $XY \in adj_{\mathcal{C}_2}$ *with* $|pds_{\mathcal{C}_2}(X, \bullet) \setminus \{Y\}| \geqslant \ell$ **do**
      while $\exists\, \mathbf{Y} \subset pds_{\mathcal{C}_2}(X, \bullet) \setminus \{Y\}$ *not yet considered with* $|\mathbf{Y}| = \ell$ **do**
        if $\mathrm{Indep}(X; Y | \{\mathbf{Y} \cup \mathbf{S}\})$ *at conf. level* α **then**
          XY **link is** non-essential and **removed**
          separation set of XY: $\mathrm{Sep}_{XY} = \{U_i\}$
        end
      end
    end
    Set $\ell = \ell + 1$
  end
end
Set all edges in $\mathcal{C}_2$ as $\circ\!\!-\!\!\circ$

4. Final orientation, out: $\mathcal{C}_4$

forall the *unshielded triples in* $\mathcal{C}_4$ **do**
  R_0: $\{X *\!\!-\!\!\circ Z \circ\!\!-\!\!* Y \;\&\; X *\!\!\not-\!\!* Y \;\&\; Z \notin \mathrm{Sep}_{XY}\} \;\Rightarrow \{X *\!\!\rightarrow Z \leftarrow\!\!* Y\}$
end

5. Propagation, out: $\mathcal{C}$

repeat
  Rules R_1 to R_{10} from [Zhang, 2008a]
until *no further orientation can be propagated*;

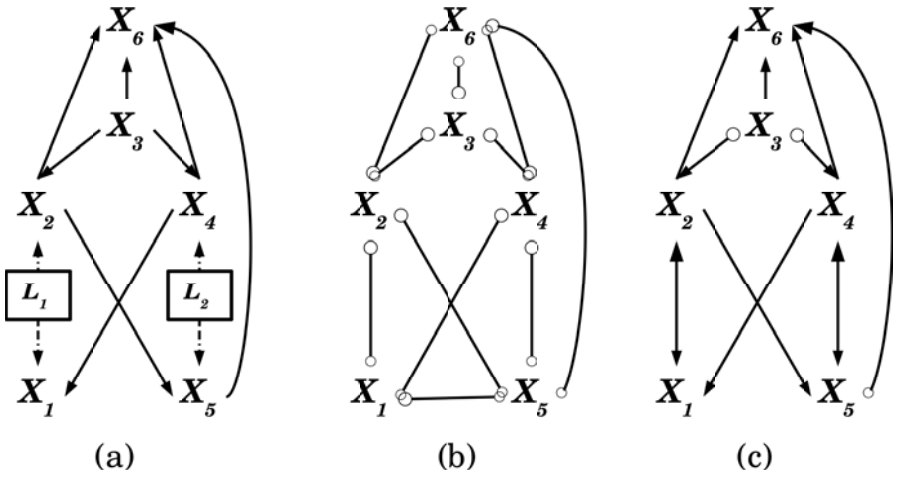


FIGURE 3.1 **DAG $\mathcal{G}$ over $\mathbf{V} = \mathbf{O} \cup \mathbf{L} \cup \mathbf{S}$ and corresponding skeleton and PAG.** $\{X_i\} \subseteq \mathbf{O}$ where $\mathbf{O}$ is the set of observed variables and $\{L_i\} \subseteq \mathbf{L}$ where $\mathbf{L}$ is the set of latent variables. The PAG (c) corresponding to the underlying DAG (a) has been reconstructed using the FCI algorithm and the conditonal independence from observed variables (see main text).

$\{X_2, X_3, X_4\}$ and $pds_{\mathcal{C}_2}(X_5, X_1) \setminus \{X_1, X_5\} = \{X_2, X_3, X_4, X_6\}$, the FCI process can identify the conditional independence.

The FCI algorithm appears rather convoluted and is known to be time consuming, as the Possible-D-SEP can get very large, even when considering sparse graphs.

By contrast, the 3off2 algorithm can easily deal with the assumption of latent variables by allowing the inclusion of observed variables that are not direct neighbours of the pair of nodes for which the conditional independence is tested. As detailed in section 6.4.2, the 3off2 requires only simple modifications that still guarantee the discovery of a causal graphical model in a reasonable amount of time on sparse graphs.

3.2.2.2 Improvements of the FCI algorithm

The RFCI [Colombo et al., 2012] and the AnytimeFCI [Spirtes, 2001] algorithms improve the speed of the FCI algorithm that suffer from an exponential time complexity due to the potentially large size of the Possible-D-SEP sets of variables.

Specifically, the RFCI approach follows the same three steps as the PC algorithm, but with modifications in the steps 2 and 3 to take into account the hidden variables. In fact, in the step 2, the RFCI algorithm remove extra edges before orienting the v-structures ([Colombo et al., 2012], Lemma 3.1) and then used a modified version of the rule R_4 for the orientation of the *discriminating paths*

([Colombo et al., 2012], Lemma 3.2). Avoiding to compute all the subsets of the Possible-D-SEP reduce the time complexity, yet the RFCI algorithm produces a less informative output than with the FCI approach. By contrast, the AnytimeFCI requires an additional adjustable parameter to perform fewer conditional tests as the FCI procedure by limiting the size of the conditioning set of variables.

As stipulated earlier, the 3off2 method combines constraint-based and Bayesian inference methods. Chapter 4 provides an overview on Bayesian structure discovery approaches that follow the *search-and-score* paradigm. Bayesian approaches have the advantage to not rely on statistical tests that are prone to erroneous removal on finite datasets. The search-and-score approaches also provide a global score for the learn causal graphical model, allowing the comparison between different structure candidates. Yet, classical Bayesian inference approaches need to exclude the possibility of hidden latent variables.

Chapter 4

Bayesian methods

The Bayesian methods [Cooper and Herskovits, 1992] step through a space of possible causal structures, usually the space of DAGs, following a *search-and-score* procedure that aims at optimizing a scoring function. This score provides an indicator of the fitness between the model and the data and thus explores each graphical model in a quantitative manner. There exist various search strategies (*e.g.* simulated annealing [Chickering et al., 1995], Markov Chain Monte Carlo [Kocka and Castelo, 2001], ant colony optimization [de Campos and Huete, 2002]) and types of score (*e.g.* entropy-based [Chow and Liu, 1968, de Campos, 2006], Minimum Description Length [Bouckaert, 1993, Rissanen, 1978], Bayesian Dirichlet score with score equivalence [Heckerman et al., 1995]).

Although the constraint-based methods are based on rigorous theoretical founding (Chapter 3), the irreversible pruning process may lead to unreliable results when applied to finite datasets. By contrast, the Bayesian approaches are more robust as search-and-score methods are typically reversible and thus do not trigger cascading errors in the following steps of the exploration procedure. However, the application of these score-based approaches to the discovery of causal graphical models remains difficult when assuming the presence of latent variables. Yet, strategies such as STRUCTURAL EM[1]-based algorithms have been proposed [Friedman, 1998, Mitchell, 1997]. Finally, background knowledge of Bayesian methods can be set, either for the causal structure discovery or for the parameters inference, in the shape of *prior* functions.

The 3off2 learning approach combines constraint-based and Bayesian frameworks to reconstruct graphical models in a robust manner despite inherent sampling noise

[1] Expectation Maximization

in finite observational datasets. This chapter provides an overview of some of the well-known scoring functions used by the Bayesian structure discovery approaches.

4.1 Learning a Bayesian Network

A Bayesian network $\mathcal{G}$ defines a joint probability distribution of a m-dimensional multivariate data vector $\mathbf{X} = (X_1, ..., X_m)$ [Pearl, 1988] and possesses two components: (i) a causal structure that entails the (in)dependence relationships between the observed variables, and (ii) a set of probability distributions that quantify these relationships. We are in particular interested in learning the first component, that can be represented by a DAG (section 2.1). Learning the structure of a Bayesian network consists in finding the DAG that best fits the data according to a given scoring function, in other words, it consists in finding the graphical model $\mathcal{G}$ that optimizes the *posterior probability* $P(\mathcal{G}|\mathcal{D})$, where $\mathcal{D}$ is the dataset with N i.i.d. samples of the vector $\mathbf{X}$. The posterior probability is related to the *likelihood*, $P(\mathcal{D}|\mathcal{G})$, of the data $\mathcal{D}$ given a graphical model $\mathcal{G}$, and to the joint probability, $P(\mathcal{G}, \mathcal{D})$, through the Bayes Formula as,

$$P(\mathcal{G}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{G})P(\mathcal{G})}{P(\mathcal{D})} = \frac{P(\mathcal{G}, \mathcal{D})}{P(\mathcal{D})} \tag{4.1}$$

where $P(\mathcal{G})$ corresponds to a prior on the graphical model $\mathcal{G}$, and $P(\mathcal{D})$ is a prior on the data $\mathcal{D}$, also called *evidence*.

When solving the problem of learning a Bayesian network, the evidence being the same for all possible graphical models, maximizing the posterior probability $P(\mathcal{G}|\mathcal{D})$ is equivalent to maximize the joint probability $P(\mathcal{G}, \mathcal{D})$. Besides, when assuming a uniform prior $P(\mathcal{G})$, maximizing the posterior probability $P(\mathcal{G}|\mathcal{D})$ turns out to be also equivalent to maximizing the likelihood $P(\mathcal{D}|\mathcal{G})$. In the following section, a uniform prior $P(\mathcal{G})$ is assumed and the likelihood $P(\mathcal{D}|\mathcal{G})$ is used as a *score*.

The search-and-score procedures are typically based on a score that is *decomposable*, to allow a computationally feasible process. Some scores are also *Markov equivalent*, ensuring that they return the same score for networks in the same equivalence class.

Decomposability Searching through the space of structures in an efficient manner requires *locally decomposable* scoring functions. After a modification on a

structure $\mathcal{G}_1$ (*e.g.* an edge removal), leading to a new structure $\mathcal{G}_2$, a decomposable scoring function allows the evaluation of $\mathcal{G}_2$ without recomputing the whole score but only by taking into account the score variation due to the modification. In other words, a decomposable scoring function can be written as a sum of local scores (in the logarithmic space) that depend only on each node and its parents.

Score equivalence When searching through the space of equivalence classes of DAGs, it is preferable that the scoring function attributes the same weight to causal structures belonging to the same equivalence class. For instance, one would expect the three possible non-v-structures of an unshielded triple $\langle X_i, X_k, X_j \rangle_{X_i \neq X_j}$ (Fig. 2.1) to receive the same score, but a different score than the v-structure $X_i \rightarrow X_k \leftarrow X_j$. [Chickering et al., 1995] shown that the MDL (Eq. 4.8) score is score equivalent. By contrast, the K2, BD or BDe scores are not score equivalent.

In the sections 4.2.1 & 4.2.2, examples of most popular *decomposable* and *score equivalent* scoring functions are given (a uniform prior probability $P(\mathcal{G})$ on the graphical model is assumed).

4.2 Scoring funtions

4.2.1 Bayesian scores

The **K2** is one of the first Bayesian scoring function that has been proposed [Cooper and Herskovits, 1992]. Under several assumptions, this scoring function can be expressed as,

$$\text{Score}_{K2}(\mathcal{D}|\mathcal{G}) = \sum_{i=1}^{n} \left[\sum_{j=1}^{q_i} \left[\log \left(\frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \right) + \sum_{k=1}^{r_i} \log(N_{ijk}!) \right] \right] \tag{4.2}$$

Then, [Heckerman et al., 1995] proposed the *Bayesian Dirichlet* (BD) score, a generalization of the **K2** score,

$$\text{Score}_{BD}(\mathcal{D}|\mathcal{G}) = \sum_{i=1}^{n} \left[\sum_{j=1}^{q_i} \left[\log \left(\frac{\Gamma(\alpha_{ij})}{\Gamma(N_{ij} + \alpha_{ij})} \right) + \sum_{k=1}^{r_i} \log \left(\frac{\Gamma(N_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})} \right) \right] \right] \tag{4.3}$$

where the α_{ijk}, so-called *hyperparameters*, correspond to a *a priori* distribution of the parameters. In practice, these hyperparameters are not easily defined. [Heckerman et al., 1995] established a scoring function, the *Bayesian Dirichlet equivalent*,

by assuming the likelihood equivalence. This allows a simple computation of the α_{ijk} by introducing the *imaginary* (or *equivalent*) sample size N',

$$\alpha_{ijk} = N' \times P(X_i = x_k, pa(X_i = x_j)|\mathcal{G}_c) \tag{4.4}$$

The imaginary sample size N' defines the weight that should be assigned to the prior distribution as compared to the number of data points. Larger values of N' tend to *oversmooth* the data and add two many edges to the structure. Typical values of N' are $1, 5, 10, 20, 50$. For our evaluations on benchmark networks (Chapter 7), the default value 10 has been chosen, as it corresponds to a standard value in the literature.

Finally, [Buntine, 1991] proposed a Bayesian scoring function that only depends on N' by assuming a uniform probability for each configuration $\{X_i\} \cup pa_{\mathcal{G}(X_i)}$, *i.e.* $P(X_i = x_k, pa(X_i = x_j)|\mathcal{G}_c) = \frac{1}{r_i q_i}$, which became a common assumption. The resulting score is the *Bayesian Dirichlet equivalent uniform* score,

$$\text{Score}_{BDeu}(\mathcal{D}|\mathcal{G}) = \sum_{i=1}^{n} \left[\sum_{j=1}^{q_i} \left[\log \left(\frac{\Gamma(\frac{N'}{q_i})}{\Gamma(N_{ij} + \frac{N'}{q_i})} \right) + \sum_{k=1}^{r_i} \log \left(\frac{\Gamma(N_{ijk} + \frac{N'}{r_i q_i})}{\Gamma(\frac{N'}{r_i q_i})} \right) \right] \right] \tag{4.5}$$

As can be seen from Eq. 4.5, this Bayesian score is decomposable, as well as score equivalent.

Although the 3off2 approach does not implement the BDeu criterion, its results have been compared with the networks learnt with a hill-climbing heuristic method relying on this criterion (Chapter 7, and Appendices A & B).

4.2.2 Information-theoretic scores

The idea behind the information-theoretic scoring functions is related to the principle of *compression* that can be obtained over the data $\mathcal{D}$ with an optimal code represented by a Bayesian network $\mathcal{G}$. In fact, one can define the *information content of* $\mathcal{D}$ *by* $\mathcal{G}$, *i.e.* the size of an optimal code induced by the distribution $\mathcal{G}$ encoding $\mathcal{D}$, given $\mathcal{D}$. This information content turns out to be equal to the negative of the log-likelihood function of the data $\mathcal{D}$ given $\mathcal{G}$, *i.e.* $log(P(\mathcal{D}|\mathcal{G}))$, by using the Huffman code. Thus, minimizing the information content amounts to maximize the log-likelihood $log(P(\mathcal{D}|\mathcal{G}))$, and thus equivalently to maximize the posterior probability $P(\mathcal{G}|\mathcal{D})$.

4.2.2.1 The LL scoring function

The previous considerations lead to the definition of the LL scoring function. Let us assume that each observed variable X_i has r_i levels, and that the set of parents for each variable X_i, $pa_{\mathcal{G}}(X_i)$, has $q_i = \prod_{X_p \in pa_{\mathcal{G}}(X_p)} r_p$ values. The likelihood of the dataset $\mathcal{D}$ given a network $\mathcal{G}$ can be computed by taking the product of the row probabilities,

$$P(\mathcal{D}|\mathcal{G}) = \prod_{i=1}^{m} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}} \tag{4.6}$$

and maximizing the likelihood with the normalization constraint $\sum_k \theta_{ijk} = 1$ leads to $\theta_{ijk} = \widehat{\theta_{ijk}} = \frac{N_{ijk}}{N_{ij}}$.

By taking the log of the previous maximum likelihood, this lead to the LL scoring function, defined as,

$$LL(\mathcal{D}|\mathcal{G}) = \sum_{i=1}^{m} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log \left(\frac{N_{ijk}}{N_{ij}} \right) \tag{4.7}$$

However, the LL score alone tends to *overfit* the data by favoring the reconstruction of complete networks. A common way of limiting the overfitting phenomenon is to *penalize* this score by taking into account the complexity of the graphical models, as detailed in the following section.

4.2.2.2 The MDL/BIC criterion

The *Minimum Description Length* (MDL) (or the *Bayesian Information Criterion* (BIC)) [Hansen and Yu, 2001, Rissanen, 1978] scoring function limits the overfitting phenomenon by penalizing the LL score with the number of parameters required by the network under reconstruction, thus preferring simpler Bayesian networks. Using the MDL criterion can be interpreted has the minimization of (i) the coding length of the model, plus (ii) the coding length of the representation of the data when encoded by this model. The scoring function is thus defined as,

$$\text{MDL}(\mathcal{D}|\mathcal{G}) = LL(\mathcal{D}|\mathcal{G}) - \frac{1}{2} log(N)|\mathcal{G}| \tag{4.8}$$

where $|\mathcal{G}| = \sum_{i=1}^{m}(r_i - 1)q_i$ is the network complexity, *i.e.* the number of parameters needed for the model G to encode $\mathcal{D}$. The first term of Eq. 4.8 stands for the number of bits needed to describe $\mathcal{D}$ while the second term measures the length of describing $\mathcal{G}$, where each parameter of $\mathcal{G}$ counts for $\frac{1}{2}log(N)$ bits. The MDL scoring function thus provides an appropriate *trade-off* between *precision* (maximization of the LL function) and *complexity* (minimization of the number of parameters).

As can be understood from Eq. 4.8, the MDL scoring function is locally decomposable,

$$\text{MDL}(\mathcal{D}|\mathcal{G}) = N \sum_{i=1}^{m} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \frac{N_{ijk}}{N} \log\left(\frac{N_{ijk}}{N_{ij}}\right) - \sum_{i=1}^{m} \frac{1}{2} \log(N)(r_i - 1)q_i \qquad (4.9)$$

In the context of the information-theoretic scoring functions, the quality formula of a Bayesian network can be generalized by $\phi(\mathcal{D}|\mathcal{G}) = LL(\mathcal{D}|\mathcal{G}) - f(N)|\mathcal{G}|$, where $f(N)$ is a non negative penalization. The AIC (*Akaike Information Criterion*) corresponds to the case of $f(N) = 1$.

The MDL criterion has been implemented in our hybrid approach 3off2 in order to score and compare network candidates while removing extra edges. However, the 3off2 algorithm slightly reformulated the MDL criterion as the parent nodes are replaced by an unknown separation set U_i of *upstream* nodes to be learnt simultaneously with the missing edge candidate (see section 6.2.2 for further details on the MDL criterion).

4.2.2.3 The NML criterion

As detailed in section 6.2.2, the MDL criterion tends to overestimate the encoding of parameters given the limited amount of data, and hence to increase the number of false negative edges, in particular when the number of levels r_i is high. The *Normalized Maximum Likelihodd* (NML) [Rissanen and Tabus, 2005, Shtarkov, 1987] criterion, based on the MDL principle, reduces this bias. Instead of overestimating the encoding of the required parameters, the NML uses the actual combinations of values of variables that exist in the available dataset $\mathcal{D}$. For a network $\mathcal{G}$, the normalized maximum likelihood score is defined as,

$$\text{NML}(\mathcal{D}|\mathcal{G}) = LL(\mathcal{D}|\mathcal{G}) - \mathcal{C}_N(\mathcal{G}) \qquad (4.10)$$

where $C_N(\mathcal{G})$ involves an exponential sum over all possible datasets of size N that can be approximated by the following formula [Roos et al., 2008b],

$$\mathcal{C}_{\mathcal{N}}(\mathcal{G}) \simeq f\mathcal{C}_{\mathcal{N}}(\mathcal{G}) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \log(\mathcal{C}_{N_{ij}}^{r_i}) \tag{4.11}$$

[Kontkanen and Myllymäki, 2007] proposed a linear-time algorithm for computing the $\log(\mathcal{C}_{N_{ij}}^{r_i})$ complexity in the case of N_{ij} observations of a single multinomial random variables (Algorithm 3).

Algorithm 3: Multinomial parametric complexity

Compute $log(\mathcal{C}_{N_{ij}}^{r_i})$ where $\mathcal{C}_m^l \equiv \mathcal{C}(l,m)$ is obtained by the following recurrence:

1. $C(1,m) = 1$

2. $C(2,m) = \sum\limits_{h_1+h_2=m} \frac{m!}{h_1!h_2!} \left(\frac{h_1}{m}\right)^{h_1} \left(\frac{h_2}{m}\right)^{h_2}$

3. if $(l > 2)$ then $\mathcal{C}(l,m) = \mathcal{C}(l-1,m) + \frac{m}{l-2}\mathcal{C}(l-2,m)$

Similarly as for the MDL criterion, the factorized NML scoring function is decomposable as,

$$f\text{NML}(\mathcal{D}|\mathcal{G}) = \sum_{i=1}^{m} \sum_{j=1}^{q_i} \left(\sum_{k=1}^{r_i} N_{ijk} \log \left(\frac{N_{ijk}}{N_{ij}}) - \log(\mathcal{C}_{N_{ij}}^{r_i}) \right) \right) \tag{4.12}$$

The factorized NML criterion has also been implemented in the 3off2 hybrid approach in order to score and compare network candidates before and after removing one extra edge. This criterion typically gives better results than the MDL criterion on benchmark networks for small datasets, *i.e.* when the bias due to the overestimation of the encoding of parameters is stronger (Chapter 7). Similarly as for the MDL criterion, the NML implemented in the 3off2 algorithm replaces the parent nodes with an unknown separation set $\{U_i\}$ of upstream nodes to be learnt simultaneously with the missing edge candidate (see section 6.2.2 for further details on the MDL criterion).

4.3 The search-and-score paradigm

4.3.1 The exact discovery

A simple idea when searching through the space of possible DAGs is to compute the score of all possible structures to find the Bayesian network that best fits the data. Yet, this is computationally infeasible for large networks as the number of causal graphical models grows superexponentially with the number of variables. Even when the maximum in-degree is limited *a priori*, finding an optimal Bayesian network remains NP-hard [Chickering et al., 1995].

Early developments led to algorithms that can solve the structure learning problem exactly for restricted classes of Bayesian networks, for instance if an ordering of the variables is given as input [Cooper and Herskovits, 1992]. More recent studies have proposed algorithms that compute the exact posterior probability of network structures for a moderate number of variables (less than 30) when additional constraints are enforced on the space of structures [Koivisto and Sood, 2004, Malone et al., 2011, Silander and Myllymaki, 2006].

Still, the limitations of the possible improvements of exact methods have prompted the developments of a variety of *heuristic* approaches.

4.3.2 The heuristic search approaches

Heuristic methods step through the space of possible DAGs, but do not explore the whole space exhaustively. The main differences between these searching methods stem from the local operators that progressively modify the causal structure to optimize the global score (*e.g.* edge addition, removal or inversion [Chickering et al., 1995]). These *greedy search* approaches are computationally feasible thanks to the decomposability of the scoring function.

However, the greedy search procedure is known to lead to local optima. The hill-climbing algorithm with random restarts limits the risk of being trapped in a single local optimum by performing several greedy searches from different seed networks. Other possible approaches are the simulated annealing procedures [Kirkpatrick et al., 1983], the genetic algorithms [Larrañaga et al., 1996]. As exemplified in the following section, it is possible to combine approaches in order to decide for the initial network.

The 3off2 reconstruction method uses instead a *best-first search* approach in order to identify the most likely dispensible edges. The iterative identification of these candidate edges is done thanks to a rank based on 2-point and 3-point information terms, allowing for a statistically robust search strategy.

4.4 Hybrid approaches

Hybrid methods have attempted to exploit the best of the constraint-based (Chapter 3) and Bayesian methods by combining the robustness of Bayesian scores with the attractive conceptual features of constraint-based approaches [Cano et al., 2008, Claassen and Heskes, 2012, Dash and Druzdzel, 1999, Singh and Valtorta, 1993, Tsamardinos et al., 2006]. In particular, [Dash and Druzdzel, 1999] have proposed to exploit an intrinsic weakness of the PC algorithm, its sensitivity to the order in which conditional independencies are tested on finite data (section 3.2.1), to rank these different order-dependent PC predictions with Bayesian scores.

In fact, many of the hybrid approaches proceed in two steps: the first step relies on a constraint-based procedure that produces a graphical model, which is then given as input to the second step that relies on a Bayesian method. For instance, [Singh and Valtorta, 1993] developed an algorithm that first gets an absolute temporal ordering on the variables and then uses this ordering as input to the K2 algorithm [Cooper and Herskovits, 1992]. The *Max-Min Hill-Climbing* (MMHC) algorithm [Tsamardinos et al., 2006] is a prominent example of hybrid method. The main idea is to restrict the search space using first the Max-Min Parents and Children (MMPC) constraint-based algorithm [Tsamardinos et al., 2003]. The MMPC step provides for each variable X the set of its parents and children, $\mathbf{PC}_X$. Then, the MMHC method follows a greedy hill-climbing strategy with operators *add-edge*, *delete-edge* and *reverse-edge*. The *add-edge* operator is only considered when it is consistent with the previously discovered parent and children set (*i.e.*, try $Y \rightarrow X$ if $Y \in \mathbf{PC}_X$). The use of the operator *delete-edge* within the hill-climbing search allows for the removal of edges during the orientation phase of MMHC.

More recently [Claassen and Heskes, 2012] have also combined constraint-based and Bayesian approaches to improve the reliability of conditional independencies by summing the likelihoods over compatible graphs.

By contrast, the 3off2 algorithm uses Bayesian scores to progressively uncover the best supported conditional independencies, by iteratively "taking off" the most likely indirect contributions of conditional 3-point information from every 2-point (mutual) information of the causal graph (section 6.3). In addition, using likelihood ratios (Eqs. 6.17 & 6.18) instead of likelihood sums [Claassen and Heskes, 2012] circumvents the need to score conditional independencies over a potentially untractable number of compatible graphs.

Chapter 5

Information-theoretic methods

The information-theoretic methods typically decide for the statistical independence between two variables using information-theoretic measures [Butte and Kohane, 2000, Chow and Liu, 1968, Margolin et al., 2006]. These methods have already given promising results, in particular in the field of systems biology, for large networks up to thousands variables. An important concept of information theory [Shannon, 1948] is the 2-point (mutual) information that provides a measure of dependencies between variables. In particular, while the correlation coefficients only measure *linear* (Pearson) or *monotonic* (Spearman) dependencies, the 2-point (mutual) information is also sensitive to diverse types of non-linear relationships. Hence, the 2-point information can be interpreted as a *general* measure of dependencies, which may be of great interest in the context of complex biological systems [Steuer et al., 2002].

In this chapter, we first introduce the entropy, and the 2-point and 3-point information quantities. These information-theoretic measures are at the root of our new causal structure learning method, namely 3off2, detailed in Chapter 6. The following sections are dedicated to state-of-art information-theoretic methods and hybrid methods relying on information-theoretic scores.

5.1 Information-theoretic measures

5.1.1 The Shannon entropy

[Fraser and Swinney, 1986] defined the entropy in a general way as the '*quantity of surprise you should feel upon reading the result of a measurement*'. Formally, the

entropy $H(X)$ of a random discrete variable X with r_x levels is defined as, [Cover and Thomas, 2006, Shannon, 1948]:

$$H(X) = -\sum_{i=1}^{r_x} p(x_i) \log(p(x_i)) \tag{5.1}$$

where x_i is one of the state of the variable X, and $p(x_i)$ is the probability of observing x_i. Thus, when the state of the variable X is fully determined, the entropy is null, while it is maximal when all the states have the same probability of occurrence (in the following, the log in Eq. 5.1 refers to the natural logarithm).

Similarly, the joint entropy for two discrete variables, X and Y, is defined as,

$$H(X,Y) = -\sum_{i=1}^{r_x}\sum_{j=1}^{r_y} p(x_i, y_j) \log(p(x_i, y_j)) \tag{5.2}$$

where Y has r_y levels, y_j is one of the state of the variable Y and $p(x_i, y_j)$ is the joint probability that X takes the value x_i and Y takes the value y_j.

A general way to decompose the entropy uses the conditional entropy,

$$H(X,Y) = H(X|Y) + H(Y) \tag{5.3}$$

where,

$$H(X|Y) = -\sum_{i=1}^{r_x}\sum_{j=1}^{r_y} p(x_i, y_j) \log(p(x_i|y_j)) \tag{5.4}$$

Eqs. 5.1, 5.4 and 5.3 lead to the 2-point information or *mutual information*, that can be interpreted as the reduction of the uncertainty on X and Y by the joint uncertainty of these two variables:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \geq 0 \tag{5.5}$$

Similarly, one can define the conditional 2-point information as,

$$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z) \geq 0 \tag{5.6}$$

5.1.2 The data processing inequality

The Data Processing Inequality is a well-known inequality of the information theory. It can be intuitively understood as the fact that no information processing by a downstream variable Z can enhance or restore information imperfectly transmitted from X to Y. This inequality can be formally defined as,

Theorem 5.1.1. Data Processing Inequality
Let's suppose that X, Y and Z are three variables related by the causal graphical model $X \to Y \to Z$, then $I(X;Z) \leq min(I(X;Y), I(Y;Z))$.

In other words, the variable Z cannot have more information about the variable X than the variable Y has about the variable X.

As detailed in the section 5.2.3, the ARACNe reconstruction method [Margolin et al., 2006] relies on the contrapositive of the DPI in order to remove superfluous edges during the structure discovery.

> The 3off2 algorithm also includes in its edge removal procedure a rank that is consistent with the DPI (Eq. 6.32).

5.1.3 The multivariate information

When taking into account more than two variables, it is also possible to define a multivariate information quantity [McGill, 1954]. In particular, we are interested in the 3-point information, which can be defined by the difference between the 2-point and the conditional 2-point information when introducing a third variable Z,

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z) \tag{5.7}$$

or, more generally,

$$I(X;Y;Z|\{U_i\}) = I(X;Y|\{U_i\}) - I(X;Y|\{U_i\}, Z) \tag{5.8}$$

with $\{U_i\}$, a set of variables that does not contain X, Y nor Z.

The conditional 3-point information $I(X;Y;Z|\{U_i\})$ is in fact symmetric in X, Y and Z,

$$\begin{aligned} I(X;Y;Z|\{U_i\}) &= I(X;Y|\{U_i\}) - I(X;Y|\{U_i\},Z) \\ &= I(X;Z|\{U_i\}) - I(X;Z|\{U_i\},Y) \\ &= I(Y;Z|\{U_i\}) - I(Y;Z|\{U_i\},X) \end{aligned}$$

This symmetry can be shown when decomposing the 3-point information using entropy terms,

$$\begin{aligned} I(X;Y;Z) = H(X) + H(Y) + H(Z) - H(X,Y) \\ - H(X,Z) - H(Y,Z) + H(X,Y,Z) \end{aligned}$$

It should be noted that, as opposed to the 2-point information, which is always positive, the 3-point information can also be negative. In the context of network discovery, as Eq. 5.8 is always valid, one can understand that no assumption on the underlying graphical model or on the amount of data available to estimate conditional 2-point and 3-point information terms is necessary.

Chapter 6 details how the 3off2 algorithm is taking advantage of Eq. 5.8, which is used to prove Lemmas 1, 3 & 4, as well as Proposition 5 (section 6.1.3). This allows to trace back the origin of necessary causal relationships in a graphical model to the existence of a negative conditional 3-point information between *three* variables $\{X,Y,Z\}$, $I(X;Y;Z|\{U_i\}) < 0$, where $\{U_i\}$ accounts for a structural independency between two of them, *e.g.* $I(X;Y|\{U_i\}) = 0$ (Lemma 4, section 6.1.3).

5.1.4 The cross-entropy of a Bayesian network

The maximum likelihood $\mathcal{L}_{\mathcal{D}|\mathcal{G}}$ of a graphical model $\mathcal{G}$ over a set of variables $\{X_i\}$ (section 4.1) is related to the cross-entropy $H(\mathcal{G},\mathcal{D}) = -\sum_{\{X_i\}} p(\{X_i\}) \log(q(\{X_i\}))$ between the "true" probability distribution $p(\{X_i\})$ from the data $\mathcal{D}$ and the approximate probability distribution $q(\{X_i\}) = \prod_i p(X_i|\{Pa_{X_i}\})$ generated by the Bayesian network $\mathcal{G}$ with specific parent nodes $\{Pa_{X_i}\}$ for each node X_i, leading to [Sanov, 1957],

$$\mathcal{L}_{\mathcal{D}|\mathcal{G}} = e^{-NH(\mathcal{G},\mathcal{D})} = e^{-N\sum_i H(X_i|\{Pa_{X_i}\})} \tag{5.9}$$

where $\sum_i H(X_i|\{Pa_{X_i}\})$ is the (conditional) entropy of the underlying causal graph.

This enables to score and compare alternative models through their maximum likelihood ratio as,

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}'}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}}} = e^{-N\sum_i \left(H(X_i|\{Pa'_{X_i}\})-H(X_i|\{Pa_{X_i}\})\right)} \tag{5.10}$$

Note, in particular, that the significance level of the Maximum likelihood approach is set by the number N of independent observational data points.

5.2 State-of-the-art of information-theoretic methods

5.2.1 The Chow & Liu approach

The Chow & Liu approach [Chow and Liu, 1968] has been the first structure learning method relying on the 2-point information measure. The output is a Maximum Spanning Tree (MST), obtained with the Kruskal's algorithm, where each edge is weighted by a 2-point information measure. The main drawback of this method is that it outputs a sparse model although the underlying networks might not be sparse.

5.2.2 Relevance Network

The *Relevance Network* (RELNET) approach [Butte and Kohane, 2000], primarily designed for clustering, computes the 2-point information of each pair of variables, and removes the edges of the pairs that have a mutual information lower than a given threshold. This method is computational efficiency, however its main limitation resides in the choice of the threshold, as a high threshold increases the number of *false negative* edges, while a low threshold increases the number of *false positive* edges, and hence impedes the precision. In fact, the 2-point information between two variables X and Y may result from a common cause Z, or from *background correlations* (*e.g.* uneven samples, normalization failure), and thus indicates a dependence although no direct relationship actually exists between X and Y.

The *Context Likelihood of Relatedness* (CLR) algorithm [Faith et al., 2007] is an extension of the RELNET approach that takes into account the background

correlations by comparing the mutual information of each pair of variables to the background distribution of mutual informations of all other possible pairs between X or Y and the remaining variables. The CLR score between two variables X and Y corresponds to $w_{xy} = \sqrt{z_x^2 + z_y^2}$, where

$$z_x = max\Big(0, \frac{I(X;Y) - \mu_x}{\sigma_x}\Big) \tag{5.11}$$

with μ_x and σ_x the mean and standard deviation of the empirical distribution of the mutual information values $I(X, X_k)$ of X with all the remaining variables.

By contrast, the 3off2 approach relies on the most significantly positive conditional 3-point information measures (section 5.1.3) to identify the most likely contributors to observed relationships and decides for the removal of edges based on a criterion related to the complexity of the graphical model under reconstruction (section 6.2.2).

In [Watkinson et al., 2009], the authors substitute the mutual information $I(i;j)$ with the sum $[I(G_i;G_j) + S(i;j)]$, where $S(i;j)$ is the *synergistic regulation index*. This index measures the degree of confidence that a variable G_i *cooperatively* regulates another variable G_j with a third variable G_k, and is defined as $S(i,j) = max_k(-I(G_i;G_j;G_k))$, where $(k \neq i)$, $(k \neq j)$, $I(G_i;G_k) < I(G_i;G_j)$ and $I(G_i;G_k) < I(G_k;G_j)$. The resulting Synergy-Augmented CLR (SA-CLR) algorithm [Watkinson et al., 2009] aims at ascertaining the existence of an interaction between G_i and G_j by identifying a *synergistic partner* G_k to G_i in the regulation of G_j.

While the SA-CLR approach aims at maximizing the 2-point information and minimizing the *synergy* interaction (*i.e.* $I(G_i;G_j;G_k)$ most negative), the 3off2 method relies on a score that maximizes the conditional 3-point information $I(X;Y;Z|\{U_i\})$ (likely non v-structure indicating the presence of a contributor), and minimizes the conditional 2-point information $I(X;Y|\{U_i\} \cup \{Z\})$ (most likely base-edge of the non v-structure), where $\{U_i\}$ is a set of variables that possibly *separates* the pair (X,Y) at end.

5.2.3 ARACNe

The *Algorithm for the Reconstruction of Accurate Cellular Network* (ARACNe) [Margolin et al., 2006] has been widely applied in the field of reverse engineering of regulatory networks. It is an efficient method, taking advantage of the 2-point information (Eq. 5.5) and of the DPI (Eq. 5.1.1). As a first step, the 2-point information, $I(X;Y)$, of all pairs of variables are computed and compared to a given threshold I_0. Each pair having a 2-point information less than I_0 is removed. Then, relying on the DPI, the ARACNe method removes also the edge of all closed triples corresponding to the weakest 2-point information. Indeed, the ARACNe methods is *assuming* that each closed triples is embedding a non-v-structure and thus that the weakest edge indicates an indirect relationship that should be discarded from the graphical model.

Although the ARACNe algorithm has shown satisfactory reconstructions on large networks, the systematic opening of the closed triples prevents from finding possible cycles and tends to increase the number of false negative edges. A parameter ϵ has been proposed to *smooth* the systematic removal of the weakest edge of certain closed triples. This ϵ corresponds to a percent tolerance threshold. Thus, if the weakest edge of the close triple $\langle X, Z, Y\rangle$ is $X - Z$, it will be removed only if $I(X;Z) \leq I(X;Y)(1-\epsilon)$ and $I(X;Z) \leq I(Y;Z)(1-\epsilon)$. One remaining issue is the absence of orientations as the ARACNe algorithm only provides a skeleton.

By contrast, as will be shown in section 6.1.2, the 3off2 algorithm is using the 3-point information in order to decide whether a non-v-structure is more likely than a v-structure. Then, the decision of removing an edge only takes place after a full decomposition of the 2-point information decomposition has been uncovered in terms of 3-point information contributions (section 6.3.1), allowing for a robust discovery of the causal structure.

5.2.4 Minimum Redundancy Networks

The *Minimum Redundancy Networks* (MRNET) approach [Meyer et al., 2007] is based on the *Maximum Relevance Minimum Redundancy* (MRMR) feature selection procedure [Peng et al., 2005]. For a given *target* variable X_j, the MRMR approach ranks a set of selected feature variables $\mathbf{X}_{\mathbf{s_j}} \subseteq \mathbf{X} \setminus X_j$ according to a score that is the difference between:

- the *maximum relevance*, represented by the 2-point information between X_j and the feature candidate X_i
- the *minimum redundancy*, represented by the average 2-point information between the feature candidate X_i and the already selected features in $\mathbf{X_{s_j}}$

This score is expected to rank well the feature variable having direct interactions with X_j. The MRNET approach repeats the MRMR greedy search, taking successively all the variables as target X_j. Formally, given a set $\mathbf{S_j}$ of already selected features for the variable X_j, $\mathbf{S_j}$ is updated by choosing X_i that maximizes the score s_i:

$$s_i = I(X_i; X_j) - \frac{1}{|\mathbf{S_j}|} \sum_{k \neq i} I(X_i; X_k) \tag{5.12}$$

Hence, for each X_j, the approach iteratively builds the set $\mathbf{X_{s_j}}$. This iterative process stops when $\frac{1}{|\mathbf{S_j}|} \sum_{k \neq i} I(X_i; X_k) > I(X_i; X_j)$. Finally, for each pair of variables (X_i, X_j), the MRNET approach computes two scores, s_i and s_j (Eq. 5.12). If the maximum of these two scores is below a given threshold, the edge $X_i - X_j$ is removed.

The limitation of this efficient method is related to an issue shared by the forward selection approaches, namely the strong dependence of the quality of the set $\mathbf{X_{s_j}}$ to the choice of the first selected feature variable, as an erroneous first feature variable impedes the following redundancy minimization process. [Meyer et al., 2010] proposed the *MRNET Backward* (MRNETB) method to overcome this limitation. As can be understood from Eq. 5.12, the set $\mathbf{X_{s_j}}$ maximizes the difference between the relevance (Eq. 5.14) of all selected variables and their redundancy (Eq. 5.15) as,

$$\mathbf{X_{s_j}} = argmax_{\mathbf{S_j}}(u_r) \tag{5.13}$$

$$u = \frac{1}{|\mathbf{S_j}|} \sum_{i \in \mathbf{S_j}} I(X_i; X_j) \tag{5.14}$$

$$r = \frac{2}{|\mathbf{S_j}|(|\mathbf{S_j}| - 1)} \sum_{i,k > i \in \mathbf{S_j}} I(X_i; X_k) \tag{5.15}$$

The MRNETB method uses a background removal procedure that starts from sets $\mathbf{X_{s_j}}$ containing all the variables and that iteratively removes the variables X_i which induce the highest increase of the $\mathbf{X_{s_j}}$ score (Eq. 5.13).

By contrast, the 3off2 reconstruction method selects the variables belonging to the set of upstream nodes $\{U_i\}$ following a rank that indicates the most likely contributors to the observed relationships (section 6.32). This rank, based on 2-point and 3-point information terms, allows for a robust discovery of conditional independencies and leads to the progressive recovery of the separation set of each pair of variables. Besides, similarly as for ARACNe, the MRNET and MRNETB approaches relies on the symmetric 2-point information measure, and thus cannot learn oriented networks. By associated the 3-point information to the 2-point information, the 3off2 algorithm can by contrast discover an oriented causal structure (see Lemma 4. in section 6.1.3 and Eq. 6.32).

In [Bontempi and Meyer, 2010], the authors use the *synergy* measure (section 5.2.2) in a feature selection procedure named min-Interaction Max-Relevancy (mIMR) to learn graphical models. The forward step of this approach consists in selecting the variables x_k that maximize the 2-point information with the target y, $I(x_k; y)$ (Max-Relevance), and minimize the multivariate information, $I(\mathbf{X}_s; x_k; y)$ (min-Interaction), where $\mathbf{X}_s$ is the set of already selected variables. Hence, the incremental step of mIMR follows $x^*_{d+1} = argmax_{x_k \in \mathbf{X} \setminus \mathbf{X}_s}(I(x_k; y) - I(\mathbf{X}_s; x_k; y))$, where x^*_{d+1} is the optimal variable to be inserted in the set $\mathbf{X}_s$ containing already d selected variables.

Instead of following a feature selection approach, the 3off2 method relies on a constraint-based procedure. Indeed, 3off2 identifies conditional independences between pairs of variables by iteratively collecting the most likely contributors, $\{U_i\}$, to an edge, $X - Y$. This identification relies on a score that maximizes the conditional 3-point information $I(X; Y; Z|\{U_i\})$ (identification of most likely non v-structures), and minimizes the conditional 2-point information $I(X; Y|\{U_i\} \cup \{Z\})$ (identification of the most likely base-edge), where $\{U_i\} \cup \{Z\}$ is a set of variables that possibly *separates* the pair (X, Y) at end.

5.2.5 The MI-CMI algorithm

To learn the structure of genetic regulatory networks, [Liang and Wang, 2008] have proposed to rely on both the 2-point and conditional 2-point information in order to evaluate the dependence between variables. The corresponding *Mutual Information-Conditional Mutual Information* (MI-CMI) algorithm first uses the mutual information to find most of the direct relationships among the variables.

The removal of edges relies on a given mutual information threshold, I_{th}. Then, the conditional mutual information is computed either for fully connected or fully unconnected triples in order to detect indirect regulations, or interactive regulations. Given the threshold I_{th} and the conditional mutual information value, edges can then be added or deleted (see [Liang and Wang, 2008], Algorithm 3).

In [Zhao et al., 2008], the authors have proposed the Direct Connectivity Metric (DCM) [Zhao et al., 2008], based on the 2-point information, $I(X;Y)$, and the conditional 2-point information, $I(X;Y|Z)$, given a third variable Z. More precisely, the DCM corresponds to a score that simultaneously maximizes the 2-point and the least conditional 2-point information, as,

$$DCM(X;Y) = \eta(X;Y) = g\left(I(X;Y), min_{Z\in\mathbf{V}\setminus\{X,Y\}} I(X;Y|Z)\right) \tag{5.16}$$

Thus, higher DCM values indicate that two variables are more likely to be directly related. Based on this metric, Zhao *et al.* devised an algorithm (see [Zhao et al., 2008], Simplified Algorithm) that ranks all pairs of variables using the DCM and, given a specified number of edges e, outputs an undirected network.

By contrast with these two methods, the 3off2 approach successively recomputes the conditional 2-point information of each pair of variables (X, Y) by iteratively building a set of variables $\{U_i\}$ that possibly *separates* (X, Y). This iterative scheme relies on the simultaneous maximization of the 3-point information conditioned on the $\{U_i\}$ set of variables and the minimization of the 2-point information also conditioned on the $\{U_i\}$ set. This scheme allows for a robust discovery of the underlying causal graphical model, without the need for input parameters, such as the number of expected edges or (conditional) 2-point information thresholds.

5.3 Hybrid approaches using *Mutual Information Tests*

This section presents a scoring function for learning Bayesian networks through a search-and-score process that performs statistical independence tests based on the χ^2 distribution in association with the 2-point information [de Campos, 2006].

[de Campos, 2006] proposed a scoring function based on the (mutual) 2-point information, called the *Mutual Information Tests* score, defined as:

$$MIT(\mathcal{G}|\mathcal{D}) = \sum_{\substack{i=1 \\ pa_{\mathcal{G}}(X_i)\neq\emptyset}} \left(2NI(X_i; pa_{\mathcal{G}}(X_i)) - \sum_{j=1}^{s_i} \chi_{\alpha, l_{i\sigma_i^\star(j)}}\right) \tag{5.17}$$

where $I(X_i; pa_{\mathcal{G}}(X_i))$ is the 2-point information between each node X_i and its parents. The second term of this scoring function is a penalization of the strength of the relationship between each variable and its parents to avoid adding variables that in fact do not belong to the parents set but may still increase the mutual information term.

The penalization term is related to the χ^2 test of independence, with α being the significance level and $\sigma_i^\star = \{\sigma_i^\star(1), ..., \sigma_i^\star(s_i)\}$ being any permutation of the index set $(1, ..., s_i)$ of the parents $pa_{\mathcal{G}}(X_i) = \{X_{i1}, ..., X_{is_i}\}$. As stated in [de Campos, 2006], the decomposition of the 2-point information $I(X_i; pa_{\mathcal{G}}(X_i))$, when including the parents one by one is not unique and this order impacts the value of the penalizing term. However, if the permutation of $pa_{\mathcal{G}}(X_i)$ satisfies $r_{i\sigma_i^\star(1)} \geq r_{i\sigma_i^\star(2)} \geq ... \geq r_{i\sigma_i^\star(s_i)}$[1], where r_{ij} represents the number of possible configurations when the parent set of X_i is restricted only to X_j, then it is guaranteed to produce the maximum penalization (see Theorem 2 [de Campos, 2006]). In this way, the $s_i!$ permutations are no longer necessary. The MIT score can be used within a search-and-score procedure, although it does not evolve in the space of equivalent DAGs, but in a reduced search space of Restricted Acyclic Partially Directed Graphs (RPDAGs) [Acid and de Campos, 2003].

As detailed in Chapter 6, the 3off2 algorithm is not trying to directly identifying the parents of each node by quantifying the strength of the relationships between each node and its parents candidate. Instead, it follows a constraint-based procedure and progressively builds a set of variables $\{U_i\}$ that separates pairs of variables because of their contribution to the observed relationship. This identification of non-essential edge relies on the 2-point and the 3-point information, and on previously introduced network complexity criterion, such as the MDL and the NML scores (Eq. 4.8 & 4.10).

[1] In other words, the first variable has the greatest number of levels, the second variable has the second largest number of levels, and so on.

Chapter 6

3off2: a hybrid method based on information statistics

This chapter presents a new hybrid method and its corresponding 3off2 algorithm [Affeldt and Isambert, 2015, Affeldt et al., 2015]. It is directly inspired by the PC anc IC algorithms but relies on a quantitative information theoretic framework to reliably uncover conditional independencies in finite datasets and subsequently orient and propagate edge directions between connected variables.

As detailed in Chapter 3, constraint-based approaches, such as the PC [Spirtes and Glymour, 1991] and IC [Pearl and Verma, 1991] algorithms, infer causal graphs from observational data, by searching for conditional independencies among variables. Under the Faithfulness and causal sufficiency assumptions, these algorithms return a Complete Partially Directed Acyclic Graph (CPDAG) that represents the Markov equivalent class of the underlying causal structure [Pearl, 2009, Spirtes et al., 2000] (section 2.1.2.2). However, as previously stated, the sensitivity of the constraint-based methods to the adjustable significance level α used for the conditional independence tests and to the order in which the variables are processed (Algorithm 1, step 1) favours the accumulation of errors when the search procedure relies on finite observational data.

The 3off2 algorithm aims at improving constraint-based methods by uncovering the most reliable conditional independencies supported by the (finite) available data, based on a quantitative information theoretic framework.

6.1 Information theoretic framework

As detailed in section 5.1.4, the maximum likelihood of a graphical model $\mathcal{G}$ is related to the cross entropy, $H(\mathcal{G},\mathcal{D}) = -\sum_{\{X_i\}} p(\{X_i\}) \log(q(\{X_i\}))$, between the "true" probability distribution $p(\{x_i\})$ from the data $\mathcal{D}$ and the approximate probability distribution $q(\{x_i\}) = \prod_i p(x_i|\{Pa_{x_i}\})$ generated by the Bayesian network $\mathcal{G}$. This relation can be used to distinguish v-structure graphical models from non-v-structure graphical models either in isolation (section 6.1.1) or embedded in a larger graph (section 6.1.2).

6.1.1 Inferring *isolated* v-structures *vs* non-v-structures

Applying the likelihood definition of a graphical model, Eq. 5.9, to isolated v-structures (Figure 6.1A) and Markov equivalent non-v-structures (Figure 6.1B-D), one obtains,

$$\begin{aligned} \mathcal{L}_{\mathsf{v}}(XY) &= e^{-N[H(Z|X,Y)+H(X)+H(Y)]} \\ &= e^{-N[H(X,Y,Z)+I(X;Y)]} \end{aligned} \tag{6.1}$$

where $I(X;Y) = H(X) + H(Y) - H(X,Y)$ is the 2-point mutual information between X and Y (Eq. 5.5), and,

$$\begin{aligned} \mathcal{L}_{\mathsf{nv}}(XY) &= e^{-N[H(X|Z)+H(Y|Z)+H(Z)]} \\ &= e^{-N[H(X,Y,Z)+I(X;Y|Z)]} \end{aligned}$$

where $I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z)$ is the conditional mutual information between X and Y given Z (Eq. 5.6). Hence, one obtains the likelihood ratio,

$$\frac{\mathcal{L}_{\mathsf{v}}(XY)}{\mathcal{L}_{\mathsf{nv}}(XY)} = e^{-N[I(X;Y)-I(X;Y|Z)]} = e^{-NI(X;Y;Z)} \tag{6.2}$$

involving the 3-point information, $I(X;Y;Z) = I(X;Y) - I(X;Y|Z)$, Eq. 5.7.

As long recognized in the field [Han, 1980, McGill, 1954], 3-point information, $I(X;Y;Z)$, can be positive or negative (if $I(X;Y) < I(X;Y|Z)$), unlike 2-point mutual information, which are always positive, $I(X;Y) \geqslant 0$.

More precisely, Eq. 6.2 demonstrates that the sign and magnitude of 3-point information provide a quantitative estimate of the relative likelihoods of isolated

v-structures *versus* non-v-structures, which are in fact independent of their actual non-connected bases XY, XZ or YZ,

$$\frac{\mathcal{L}_{\mathsf{v}}(XY)}{\mathcal{L}_{\mathsf{nv}}(XY)} = \frac{\mathcal{L}_{\mathsf{v}}(XZ)}{\mathcal{L}_{\mathsf{nv}}(XZ)} = \frac{\mathcal{L}_{\mathsf{v}}(YZ)}{\mathcal{L}_{\mathsf{nv}}(YZ)} = e^{-NI(X;Y;Z)} \tag{6.3}$$

Hence, a significantly negative 3-point information, $I(X;Y;Z) < 0$, implies that a v-structure is more likely than a non-v-structure given the observed correlation data. Conversely, a significantly positive 3-point information, $I(X;Y;Z) > 0$, implies that a non-v-structure model is more likely than a v-structure model.

Yet, Eq. 6.3 shows that, as the 3-point information, $I(X;Y;Z)$, is symmetric by construction, it cannot indicate how to orient v-structures or non-v-structures over the XYZ triple. To this end, it is however straightforward to show that the most likely base (XY, XZ or YZ) of the local v-structure or non-v-structure corresponds to the pair with lowest mutual information, *e.g.*, $I(X;Y) = \min_{XYZ}\big(I(S;T)\big)$, as shown by the likelihood ratios,

$$\frac{\mathcal{L}_{\mathsf{v}}(XY)}{\mathcal{L}_{\mathsf{v}}(ST)} = \frac{\mathcal{L}_{\mathsf{nv}}(XY)}{\mathcal{L}_{\mathsf{nv}}(ST)} = \frac{e^{-NI(X;Y)}}{e^{-NI(S;T)}}$$

Note, in particular, that choosing the base with the lowest mutual information is consistent with the Data Processing Inequality (Theorem 5.1.1) expected for non-v-structures, Figures 6.1B-D.

Hence, combining 3-point and 2-point information allows to determine the likelihood *and* the base of isolated v-structures *versus* non-v-structures. But how to extend such simple results to identify local v-structures and non-v-structures embedded within an entire graph $\mathcal{G}$?

6.1.2 Inferring *embedded* v-structures *vs* non-v-structures

To go from isolated to embedded v-structures and non-v-structures within a DAG $\mathcal{G}$, we will consider the markov equivalent CPDAG of $\mathcal{G}$ (section 2.1.2.2) and introduce generalized v-structures and non-v-structures, Figures 6.1E-H. We will demonstrate that their relative likelihood, given the available observational data, can be estimated from the sign and magnitude of a conditional 3-point information, $I(X;Y;Z|\{U_i\})$, Eq. 6.4. This will extend our initial result valid for isolated v-structures and non-v-structures, Eq. 6.3.

Let's consider a pair of non-neighbor nodes X, Y with a set of upstream nodes $\{U_i\}_n$, where each node U_i has at least one direct connection to X ($U_i \to X$) or Y ($U_i \to Y$) or to another upstream node $U_j \in \{U_i\}_n$ ($U_i \to U_j$) or only undirected links to these nodes ($U_i - X$, $U_i - Y$ or $U_i - U_j$).

Thus, given X, Y and a set of upstream nodes $\{U_i\}_n$, any additional node Z can either be:

- *i)* at the apex of a generalized v-structure, if *all* existing connections between X, Y, $\{U_i\}_n$ and Z are directed and point towards Z, Figure 6.1E, or else,
- *ii)* Z has at least one undirected link with X, Y or one of the upstream nodes U_i or at least one directed link pointing towards these nodes ($Z \to X$, $Z \to Y$ or $Z \to U_i$), Figures 6.1F-H. In such a case, Z might contribute to the mutual information $I(X;Y)$ and should be included in the set of upstream nodes $\{U_i\}_n$, thereby defining a generalized non-v-structure, Figures 6.1F-H.

Then, similarly to the case of an isolated v-structure (Eq. 6.1), the maximum likelihood $\mathcal{L}_{\mathsf{v}}(XY)$ of a generalized v-structure pointing towards Z from a base XY with upstream nodes $\{U_i\}_n$ can be expressed as,

$$\begin{aligned}\mathcal{L}_{\mathsf{v}}(XY) &= e^{-N[H(Z|X,Y,\{U_i\})+H(X|\{U_i\})+H(Y|\{U_i\})+H(\{U_i\})]} \\ &= e^{-N[H(X,Y,Z,\{U_i\})+I(X;Y|\{U_i\})]}\end{aligned}$$

where $I(X;Y|\{U_i\})$ is the conditional mutual information between X and Y given $\{U_i\}$, $I(X;Y|\{U_i\}) = H(X|\{U_i\}) + H(Y|\{U_i\}) - H(X,Y|\{U_i\}) - H(\{U_i\})$.

Likewise, the maximum likelihood $\mathcal{L}_{\mathsf{nv}}(XY)$ of a generalized non-v-structure of base xy with upstream nodes $\{U_i\}_n$ and Z can be expressed as,

$$\begin{aligned}\mathcal{L}_{\mathsf{nv}}(XY) &= e^{-N[H(X|Z,\{U_i\})+H(Y|Z,\{U_i\})+H(Z,\{U_i\})]} \\ &= e^{-N[H(X,Y,Z,\{U_i\})+I(X;Y|Z,\{U_i\})]}\end{aligned}$$

where $I(X;Y|Z,\{U_i\}) = H(X|Z,\{U_i\})+H(Y|Z,\{U_i\})-H(X,Y|Z,\{U_i\})-H(Z,\{U_i\})$ is the conditional mutual information between X and Y given Z and $\{U_i\}$. Hence,

$$\frac{\mathcal{L}_{\mathsf{v}}(XY)}{\mathcal{L}_{\mathsf{nv}}(XY)} = e^{-NI(X;Y;Z|\{U_i\})} \tag{6.4}$$

in which the conditional 3-point information, $I(X;Y;Z|\{U_i\}) = I(X;Y|\{U_i\}) - I(X;Y|Z,\{Ui\})$, is involved (section 5.1.3).

Hence, a significantly negative conditional 3-point information, $I(X;Y;Z|\{U_i\}) < 0$, implies that a generalized v-structure is more likely than a generalized non-v-structure given the available observational data. Conversely, a significantly positive conditional 3-point information, $I(X;Y;Z|\{U_i\}) > 0$, implies that a generalized non-v-structure model is more likely than a generalized v-structure model.

Yet, as the conditional 3-point information, $I(X;Y;Z|\{U_i\})$, is in fact invariant upon permutations between X, Y and Z (section 5.1.3), it cannot indicate how to orient embedded v-structures or non-v-structures over the XYZ triple, as already noted in the case of isolated v-structures and non-v-structures, above.

However, the most likely base (XY, XZ or YZ) of the embedded v-structure or non-v-structure corresponds to the least correlated pair conditioned on $\{U_i\}$, *e.g.*, $I(X;Y|\{U_i\}) = \min_{XYZ}\big(I(S;T|\{U_i\})\big)$, as shown with the following likelihood ratios,

$$\frac{\mathcal{L}_{\mathsf{v}}(XY)}{\mathcal{L}_{\mathsf{v}}(ST)} = \frac{\mathcal{L}_{\mathsf{nv}}(XY)}{\mathcal{L}_{\mathsf{nv}}(ST)} = \frac{e^{-NI(X;Y|\{U_i\})}}{e^{-NI(S;T|\{U_i\})}} \tag{6.5}$$

Note, in particular, that choosing the base with the lowest conditional mutual information, *e.g.*, $I(X;Y|\{U_i\}) = \min_{XYZ}\big(I(S;T|\{U_i\})\big)$, is consistent with the Data Processing Inequality expected for the generalized non-v-structure of Figures 6.1F-H, $I(X;Y) \leqslant \min\big(I(X;Z,\{U_i\}), I(Z,\{U_i\};Y)\big)$, as shown below for $I(X;Y)$ and $I(X;Z,\{U_i\})$, by subtracting $I(X;Y;Z|\{U_i\})$ on each side of the inequality $I(X;Y|\{U_i\}) \leqslant I(X;Z|\{U_i\})$, leading to,

$$\begin{aligned} I(X;Y|Z,\{U_i\}) &\leqslant I(X;Z|\{U_i\},Y) \\ &\leqslant I(X;Z|\{U_i\},Y) + I(X;\{U_i\}|Y) \\ &\leqslant I(X;Z,\{U_i\}|Y) \\ I(X;Y) &\leqslant I(X;Z,\{U_i\}) \end{aligned}$$

where we have used the chain rule, $I(X;Z,\{U_i\}|Y) = I(X;Z|\{U_i\},Y) + I(X;\{U_i\}|Y)$, before adding $I(X;Y;Z,\{U_i\})$ on each side of the inequality. The corresponding inequality holds between $I(X;Y)$ and $I(Z,\{U_i\};Y)$, implying the Data Processing Inequality.

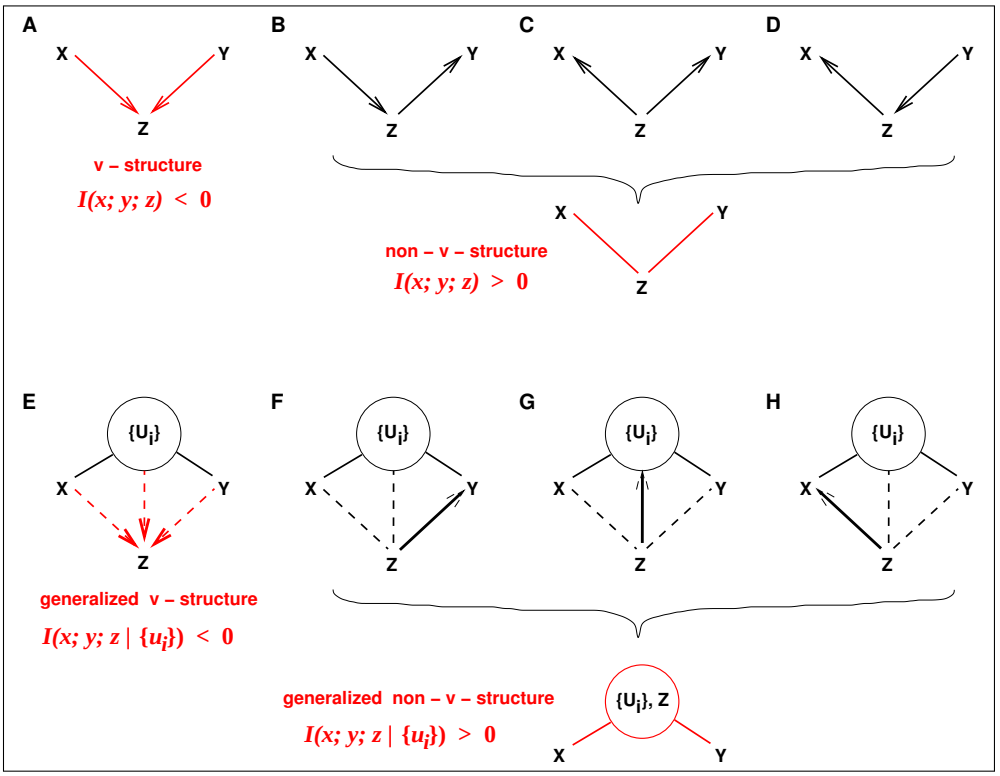


FIGURE 6.1 **Inference of v-structures versus non-v-structures by 3-point information from observational data.**

6.1.3 Uncovering causality from a stable/faithful distribution

This section formalizes the previous results on isolated and embedded v-structures and non-v-structures by establishing how the causality can be uncovered from a stable distribution, in the framework of information theory.

6.1.3.1 Negative 3-point information as evidence of causality

Consider a network $\mathcal{G} = (V, E)$ and a *stable* (or *faithful*) distribution $P(\mathbf{X})$ over V, implying that each structural independency (*i.e.* missing edge XY in $\mathcal{G}$) corresponds to a vanishing conditional 2-point (mutual) information and reciprocally as,

$$(X \perp\!\!\!\perp Y|\{U_i\})_G \iff (X \perp\!\!\!\perp Y|\{U_i\})_P \tag{6.6}$$

$$\iff I(X;Y|\{U_i\}) = 0 \tag{6.7}$$

Eq.6.6 assumes, in particular, that $P(\mathbf{X})$ is a theoretical distribution, defined by a formal expression of its variables $\mathbf{X} = \{X, Y, U_1, U_2, \ldots\}$. Note, however, that

no such expression is known *a priori*, in general, and $P(\mathbf{X})$ must typically be *estimated* from the available data. In principle, an infinite amount of data would be necessary to infer an 'exact' *stable* distribution $P(\mathbf{X})$ consistent with Eq.6.6.

In the following, we will first assume that such an infinite amount of data is available and distributed as a stable $P(\mathbf{X})$ to establish how causality can be inferred statistically from conditional 2-point and 3-point information. We will then consider the more realistic situation for which $P(\mathbf{X})$ is not known exactly and must be estimated from a finite amount of data.

The definition of the conditional 3-point information given by Eq. 5.8 provides also a generic decomposition of a conditional 2-point (X, Y) information, $I(X;Y|\{U_i\})$, by the introduction of a third node Z,

$$I(X;Y|\{U_i\}) = I(X;Y;Z|\{U_i\}) + I(X;Y|\{U_i\},Z) \tag{6.8}$$

As already stipulated in section 5.1.3, Eq.6.8 is always valid, regardless of any assumption on the underlying graphical model and of the amount of data available to estimate conditional 2-point and 3-point information terms. Eq.6.8 will be used to prove the following lemmas and propositions, which trace back the origin of necessary causal relationships in a graphical model to the existence of a negative conditional 3-point information between *three* variables $\{X, Y, Z\}$, $I(X;Y;Z|\{U_i\}) < 0$, where $\{U_i\}$ accounts for a structural independency between two of them, *e.g.* $I(X;Y|\{U_i\}) = 0$ (see Theorem 4).

Lemma 1. *Given a stable distribution* $P(\mathbf{X})$ *on* V, $\forall X, Y \in V$ *not adjacent in* $\mathcal{G}$, $\exists \{U_i\} \subseteq V_{\setminus\{X,Y\}}$ *s.t.* $I(X;Y|\{U_i\}) = 0$ *and* $\forall Z \neq X, Y, \{U_i\}$, $I(X;Y;Z|\{U_i\}) \leqslant 0$.

Proof. If $X, Y \in V$ are not adjacent in $\mathcal{G}$, this corresponds to a structural independency in $\mathcal{G}$, *i.e.* $\exists \{U_i\} \subseteq V_{\setminus\{X,Y\}}$ s.t. $I(X;Y|\{U_i\}) = 0$. Then $\forall Z \neq X, Y, \{U_i\}$ Eq.6.8 implies $I(X;Y;Z|\{U_i\}) = -I(X;Y|\{U_i\},Z) \leqslant 0$, as (conditional) mutual information is always positive. □

Corollary 2 (3-point contribution). $\forall X, Y, Z \in V$ *and* $\forall \{U_i\} \subseteq V_{\setminus\{X,Y,Z\}}$ *s.t.* $I(X;Y;Z|\{U_i\}) > 0$, *then* $I(X;Y|\{U_i\}) > 0$ *(as well as* $I(X;Z|\{U_i\}) > 0$ *and* $I(Y;Z|\{U_i\}) > 0$ *by symmetry of* $I(X;Y;Z|\{U_i\})$*).*

Corollary 2, which is a direct consequence of Equation 6.8 and the positivity of mutual information, will be the basis of the 3off2 causal network reconstruction algorithm, which iteratively "takes off" 3-point information from 2-point

information, as $I(X;Y|\{U_i\}) - I(X;Y;Z|\{U_i\}) = I(X;Y|\{U_i\},Z)$, and update $\{U_i\} \leftarrow \{U_i\} + Z$ as long as there remains some $Z \in V$ with (significantly) positive conditional 3-point information $I(X;Y;Z|\{U_i\}) > 0$.

Lemma 3 (vanishing conditional 2-point and 3-point information in undirected networks). *If $\mathcal{G}$ is an undirected (Markov) network, $\forall X, Y \in V$ and $\forall \{U_i\} \subseteq V_{\setminus\{X,Y\}}$ s.t. $I(X;Y|\{U_i\})=0$, then $\forall Z \neq X, Y, \{U_i\}$, $I(X;Y;Z|\{U_i\}) = 0$.*

Proof. If $\mathcal{G}$ is a Markov network, $\forall X, Y \in V$ and $\forall \{U_i\} \subseteq V_{\setminus\{X,Y\}}$ s.t. $I(X;Y|\{U_i\}) = 0$, then $\forall Z \neq X, Y, \{U_i\}$, $I(X;Y|\{U_i\},Z) = 0$ as conditioning observation cannot induce correlations in Markov networks [Koller and Friedman, 2009]. This implies that $I(X;Y;Z|\{U_i\}) = 0$ through Eq.6.8. □

Note, however, that the converse of Lemma 3 is not true. Namely, (partially) directed networks can also have vanishing conditional 3-point information associated to all their structural independencies. In particular, tree-like bayesian networks without colliders (*i.e.* without v-structures, $X \rightarrow Z \leftarrow Y$) present only vanishing 3-point information associated to their structural independencies, *i.e.* $I(X;Y;Z|\{U_i\}) = 0$, $\forall X, Y, Z, \{U_i\} \in V$ s.t. $I(X;Y|\{U_i\}) = 0$. However, such a directed network must be Markov equivalent to an undirected network corresponding to the same structural independencies but lacking any trace of causal relationships (*i.e.* no directed edges). The probability distributions faithful to such directed networks do not contain evidence of obligate causality; *i.e* no directed edges can be unambiguously oriented.

The following Theorem 4 establishes the existence of negative conditional 3-point information as statistical evidence of obligate causality in graphical models. For the purpose of generality in this section, we do not exclude the possibility that unobserved 'latent' variables might mediate the causal relationships among observed variables. However, this requires dissociating the labelling of the two endpoints of each edges. Let us first recall the three different endpoint marks associated to such edges in mixed graphs: they are the tail ($-$), the head ($>$) and the unspecified ($\circ$) endpoint marks. In addition, we will use the asterisk symbol ($*$) as a wild card denoting any of the three marks (see section 2.2 for further details on mixed and ancestral graphs).

Theorem 4 (negative conditional 3-point information as statistical evidence of causality). *If* $\exists X, Y, Z \in V$ *and* $\{U_i\} \subseteq V_{\setminus\{X,Y,Z\}}$ *s.t.* $I(X;Y|\{U_i\}) = 0$ *and* $I(X;Y;Z|\{U_i\}) < 0$ *then,* $\mathcal{G}$ *is (partially) directed, i.e. some variables in* $\mathcal{G}$ *are causally linked, either directly or indirectly through other variables, including possibly unknown, 'latent' variables unobserved in* $\mathcal{G}$*.*

Proof. Theorem 4 is the contrapositive of Lemma 3, with the additional use of Lemma 1. □

Proposition 5 (origin of causality at unshielded triples with negative conditional 3-point information). *for all unshielded triple,* $X \ast\!\!-\!\!\circ\ Z \circ\!\!-\!\!\ast\ Y$*,* $\exists\{U_i\} \subseteq V_{\setminus\{X,Y\}}$ *s.t.* $I(X;Y|\{U_i\}) = 0$*, if* $Z \notin \{U_i\}$ *then* $I(X;Y;Z|\{U_i\}) < 0$ *and the unshielded triple should be oriented as* $X \ast\!\!\rightarrow Z \leftarrow\!\!\ast\ Y$

Proof. if $I(X;Y|\{U_i\}) = 0$ with $Z \notin \{U_i\}$, the unshielded triple has to be a collider and $I(X;Y|\{U_i\}, Z) > 0$, by faithfulness, hence, $I(X;Y;Z|\{U_i\}) < 0$ by Eq. 6.8. □

Hence, the origin of causality manifests itself in the form of colliders or v-structures in graphical models which reveal 'genuine' causations ($X \rightarrow Z$ or $Y \rightarrow Z$) or, alternatively, 'possible' causations ($X \circ\!\!\rightarrow Z$ or $Y \circ\!\!\rightarrow Z$), provided that the corresponding correlations are not due to unobserved 'latent' variables L or L' as, $X \dashleftarrow L \dashrightarrow Z$ or $Y \dashleftarrow L' \dashrightarrow Z$.

6.1.3.2 Propagating the orientations

Following the rationale of constraint-based approaches, it is then possible to 'propagate' further the orientations downstream of colliders, through positive (conditional) 3-point information, if one assumes that the underlying distribution $P(\mathbf{X})$ is faithful to an *ancestral graph* $\mathcal{G}$ on V. An *ancestral graph* is a mixed graph, that is, with three types of edges, undirected ($-$), directed ($\leftarrow$ or $\rightarrow$) or bidirectional ($\leftrightarrow$), but with *i.)* no directed cycle, *ii.)* no almost directed cycle (including one bidirectional edge) and *iii.)* no undirected edge with incoming arrowhead (such as $X \ast\!\!\rightarrow Z - Y$). In particular, Directed Acyclic Graphs (DAG) are subclasses of ancestral graphs (*i.e.* without undirected nor bidirectional edges).

Proposition 6 ('propagation' of causality at unshielded triples with positive conditional 3-pt information). *Given a distribution $P(\mathbf{X})$ faithful to an ancestral graph $\mathcal{G}$ on V, for all unshielded triple with already one converging orientation, $X *\!\!\rightarrow Z \circ\!\!-\!\!* Y$, $\exists\{U_i\} \subset V_{\setminus\{X,Y\}}$ s.t. $I(X;Y|\{U_i\},Z) = 0$, if $Z \in \{U_i\}$ then $I(X;Y;Z|\{U_i\}) > 0$ and the first orientation should be 'propagated' to the second edge as $X *\!\!\rightarrow Z \rightarrow Y$.*

Proof. If $I(X;Y|\{U_i\}) = 0$ with $Z \in \{U_i\}$, the unshielded triple cannot be a collider and, since $\mathcal{G}$ is assumed to be an ancestral graph, the edge $Z - Y$ cannot be an undirected edge either. Hence, it has to be a directed edge, $Z \rightarrow Y$ and $I(X;Y;Z|\{U_i\}_{\setminus Z}) > 0$ by faithfulness and Eq. 6.8 □

Note that the propagation rule of Proposition 5 (*i.*) can be applied iteratively to successive unshielded triples corresponding to positive conditional 3-point information. Yet, all arrowhead orientations can be ultimately traced back to a negative conditional 3-point information, Theorem 4 and Proposition 5.

6.2 Robust reconstruction of causal graphs from finite datasets

6.2.1 Finite size corrections of maximum likelihood

We now turn to the more practically relevant situation of finite datasets consisting of N independent data points. The associated sampling noise will instrinsically limit the accuracy of causal network reconstruction. In particular, conditional independencies cannot be exactly achieved ($I(X;Y|\{U_i\}) = 0$) but can be reliably established using statistical criteria that depend on the number of data points N.

Given N independent datapoints from the available data $\mathcal{D}$, let us recall the maximum likelihood, $\mathcal{L}_{\mathcal{D}|\mathcal{G}}$, that they might have been generated by the graphical model $\mathcal{G}$ [Sanov, 1957],

$$\mathcal{L}_{\mathcal{D}|\mathcal{G}} = \frac{e^{-NH(\mathcal{G},\mathcal{D})}}{Z(\mathcal{G},\mathcal{D})} = \frac{e^{N\sum_{\{x_i\}} p(\{x_i\})\log(q(\{x_i\}))}}{Z(\mathcal{G},\mathcal{D})} \tag{6.9}$$

where $H(\mathcal{G},\mathcal{D}) = -\sum_{\{x_i\}} p(\{x_i\})\log(q(\{x_i\}))$ is the cross entropy between the "true" probability distribution $p(\{x_i\})$ of the data $\mathcal{D}$ and the theoretical probability distribution $q(\{x_i\})$ of the model $\mathcal{G}$ and $Z(\mathcal{G},\mathcal{D})$ is a data- and model-dependent factor ensuring proper normalization condition. The structural constraints of the

model $\mathcal{G}$ can be included *a priori* in the factorization form of the theoretical probability distribution, $q(\{x_i\})$.

In particular, if we assume a Bayesian network as underlying graphical model, $q(\{x_i\})$ factorizes as $q(\{x_i\}) = \prod_i p(x_i|\{\text{pa}_{x_i}\})$, where $\{\text{pa}_{x_i}\}$ denote the values of the parents of node X_i, $\{\text{Pa}_{X_i}\}$, and leads to the following maximum likelihood expression,

$$\mathcal{L}_{\mathcal{D}|\mathcal{G}} = \frac{e^{-N\sum_i H(X_i|\{\text{Pa}_{X_i}\})}}{Z(\mathcal{G},\mathcal{D})} \tag{6.10}$$

The model $\mathcal{G}$ can then be compared to the alternative model $\mathcal{G}_{\backslash X\to Y}$ with one additional missing edge $X \to Y$ using the maximum likelihood ratio,

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\backslash X\to Y}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}}} = e^{-NI(X;Y|\{\text{Pa}_Y\}_{\backslash X})}\frac{Z(\mathcal{G},\mathcal{D})}{Z(\mathcal{G}_{\backslash X\to Y},\mathcal{D})} \tag{6.11}$$

where $I(X;Y|\{\text{Pa}_Y\}_{\backslash X}) = H(Y|\{\text{Pa}_Y\}_{\backslash X}) - H(Y|\{\text{Pa}_Y\})$. However, Eq.6.11 cannot be used as such to learn the underlying graphical model, as it assumes that the order between the nodes and their parents is already known (see however [de Campos, 2006]). Yet, following the rationale of constraint-based approaches, Eq.6.11 can be reformulated by replacing the parent nodes with an unknown separation set $\{U_i\}$ to be learnt simultaneously with the missing edge candidate XY,

$$\begin{aligned}\frac{\mathcal{L}_{\mathcal{G}_{\backslash XY|\{U_i\}}}}{\mathcal{L}_{\mathcal{G}}} &= e^{-NI(X;Y|\{U_i\})+k_{X;Y|\{U_i\}}} \\ k_{X;Y|\{U_i\}} &= \log\left(Z(\mathcal{G},\mathcal{D})/Z(\mathcal{G}_{\backslash XY|\{U_i\}},\mathcal{D})\right)\end{aligned} \tag{6.12}$$

where the factor $k_{X;Y|\{U_i\}} > 0$ tends to limit the complexity of the models by favoring fewer edges. Namely, the condition, $I(X;Y|\{U_i\}) < k_{X;Y|\{U_i\}}/N$, implies that simpler models compatible with the structural independency, $X \perp\!\!\!\perp Y|\{U_i\}$, are more likely than model $\mathcal{G}$, given the finite available dataset. This replaces the 'perfect' conditional independency condition, $I(X;Y|\{U_i\}) = 0$, valid in the limit of an infinite dataset, $N \to \infty$.

A common complexity criterion in model selection is the Bayesian Information Criterion (BIC) or Minimal Description Length (MDL) criterion [Hansen and Yu, 2001, Rissanen, 1978],

$$k^{\text{MDL}}_{X;Y|\{U_i\}} = \frac{1}{2}(r_x - 1)(r_y - 1)\prod_i r_{u_i}\log N \tag{6.13}$$

where r_x, r_y and r_{u_i} are the number of levels of the corresponding variables (see also section 4.2.2.2). The MDL complexity, Eq.6.13, is simply related to the normalisation constant reached in the asymptotic limit of a large dataset $N \to \infty$ (Laplace approximation). However, this limit distribution is only reached for very large datasets in practice.

Alternatively, the normalisation of the maximum likelihood can also be done over all possible datasets including the same number of data points to yield a (universal) Normalized Maximum Likelihood (NML) criterion [Rissanen and Tabus, 2005, Shtarkov, 1987] (see also section 4.2.2.2) and its decomposable [Kontkanen and Myllymäki, 2007, Roos et al., 2008a] and XY-symmetric version, $k^{\mathsf{NML}}_{X;Y|\{U_i\}}$, defined in section 6.2.2.

Then, instead of exploring the combinatorics of sepset composition $\{U_i\}$ for each missing edge candidate XY as in traditional constraint-based approaches, we propose that Eq.6.12 can be used to iteratively extend a *likely* sepset using the maximum likelihood ratios between two successive sepset candidates, *i.e.* between the already ascertained $\{U_i\}$ and the possible extended $\{U_i\} + Z$, as,

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\},Z}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\}}}} = e^{NI(X;Y;Z|\{U_i\})+k_{X;Y;Z|\{U_i\}}} \tag{6.14}$$

using Eq.6.8 for $I(X;Y;Z|\{U_i\})$ and introducing a similar 3-point complexity conditioned on $\{U_i\}$ as,

$$k_{X;Y;Z|\{U_i\}} = k_{X;Y|\{U_i\},Z} - k_{X;Y|\{U_i\}}$$

where $k_{X;Y;Z|\{U_i\}} \geqslant 0$, unlike 3-point information, $I(X;Y;Z|\{U_i\})$ which can be positive or negative.

Introducing also the shifted 2-point and 3-point information for finite datasets as,

$$I'(X;Y|\{U_i\}) = I(X;Y|\{U_i\}) - \frac{k_{X;Y|\{U_i\}}}{N} \tag{6.15}$$

$$I'(X;Y;Z|\{U_i\}) = I(X;Y;Z|\{U_i\}) + \frac{k_{X;Y;Z|\{U_i\}}}{N} \tag{6.16}$$

leads to maximum likelihood ratios equivalent to Eqs.6.12 and 6.14,

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\}}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}}} = e^{-NI'(X;Y|\{U_i\})} \tag{6.17}$$

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\},Z}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\}}}} = e^{NI'(X;Y;Z|\{U_i\})} \tag{6.18}$$

As will become apparent in the following discussion, learning, iteratively, the most likely edge to be removed XY and its corresponding separation set $\{U_i\}$ will imply to simultaneously minimize 2-point information (Eq.6.17) while maximizing 3-point information (Eq.6.18).

6.2.2 Complexity of graphical models

This section gives further details on the complexities used by the 3off2 structure discovery approach and introduced in the section 4.2.2.

As previously stated (Eq.6.10 & 6.12), the complexity $k_{\mathcal{G},\mathcal{D}}$ of a graphical model is related to the normalization constant $Z(\mathcal{G},\mathcal{D})$ of its maximum likelihood as $k_{\mathcal{G},\mathcal{D}} = \log Z(\mathcal{G},\mathcal{D})$,

$$\mathcal{L}_{\mathcal{G}} = \frac{e^{-NH(\mathcal{G},\mathcal{D})}}{Z(\mathcal{G},\mathcal{D})} = e^{-NH(\mathcal{G},\mathcal{D})-k_{\mathcal{G},\mathcal{D}}}$$

For Bayesian networks with decomposable entropy, *i.e.* $H(\mathcal{G},\mathcal{D}) = \sum_i H(X_i|\{\mathrm{Pa}_{X_i}\})$, it is convenient to use decomposable complexities, $k_{\mathcal{G},\mathcal{D}} = \sum_i k_{X_i|\{\mathrm{Pa}_{X_i}\}}$,

$$\mathcal{L}_{\mathcal{G}} = e^{-N\sum_i H(X_i|\{\mathrm{Pa}_{X_i}\})-\sum_i k_{X_i|\{\mathrm{Pa}_{X_i}\}}}$$

such that the comparison between alternative models $\mathcal{G}$ and $\mathcal{G}_{\backslash X\to Y}$ (*i.e.* $\mathcal{G}$ with one missing edge $X \to Y$) leads to a simple local increment of the score,

$$\begin{aligned} \frac{\mathcal{L}_{\mathcal{G}_{\backslash X\to Y}}}{\mathcal{L}_{\mathcal{G}}} &= e^{-NI(X;Y|\{\mathrm{Pa}_Y\}_{\backslash X})+\Delta k_{Y|\{\mathrm{Pa}_Y\}_{\backslash X}}} \\ I(X;Y|\{\mathrm{Pa}_Y\}_{\backslash X}) &= H(Y|\{\mathrm{Pa}_Y\}_{\backslash X}) - H(Y|\{\mathrm{Pa}_Y\}) \geqslant 0 \\ \Delta k_{Y|\{\mathrm{Pa}_Y\}_{\backslash X}} &= k_{Y|\{\mathrm{Pa}_Y\}} - k_{Y|\{\mathrm{Pa}_Y\}_{\backslash X}} \geqslant 0 \end{aligned} \tag{6.19}$$

As introduced in section 4.2.2, a common complexity criterion in model selection is the Bayesian Information Criterion (BIC) or Minimal Description Length (MDL) criterion [Hansen and Yu, 2001, Rissanen, 1978],

$$k^{\mathsf{MDL}}_{Y|\{\mathrm{Pa}_Y\}} = \frac{1}{2}(r_y - 1)\prod_j^{\mathrm{Pa}_Y} r_j \ \log N \tag{6.20}$$

$$\Delta k^{\mathsf{MDL}}_{Y|\{\mathrm{Pa}_Y\}_{\backslash X}} = \frac{1}{2}(r_x - 1)(r_y - 1)\prod_j^{\mathrm{Pa}_{y_{\backslash X}}} r_j \ \log N \tag{6.21}$$

where r_x, r_y and r_j are the number of levels of each variable, x, y and j. The MDL complexity, Eq.6.20, is simply related to the normalisation constant reached in the asymptotic limit of a large dataset $N \to \infty$ (Laplace approximation). The MDL complexity can also be derived from the Stirling approximation on the Bayesian measure [Bouckaert, 1993, Schwarz, 1978]. Yet, in practice, this limit distribution is only reached for very large datasets, as some of the least-likely $(r_y - 1)\prod_j r_j$ combinations of states of variables are in fact rarely (if ever) sampled in typical finite datasets. As a result, the MDL complexity criterion tends to underestimate the relevance of edges connecting variables with many levels, r_i, leading to the removal of false negative edges.

To avoid such biases with finite datasets, the normalisation of the maximum likelihood can be done over all possible datasets with the same number N of data points. This corresponds to the (universal) Normalized Maximum Likelihood (NML, introduced in section 4.2.2) criterion [Kontkanen and Myllymäki, 2007, Rissanen and Tabus, 2005, Roos et al., 2008a, Shtarkov, 1987],

$$\mathcal{L}_{\mathcal{G}} = \frac{e^{-NH(\mathcal{G},\mathcal{D})}}{\sum_{|\mathcal{D}'|=N} e^{-NH(\mathcal{G},\mathcal{D}')}} = e^{-NH(\mathcal{G},\mathcal{D})-k^{\mathsf{NML}}_{\mathcal{G},\mathcal{D}}} \tag{6.22}$$

We introduce here the factorized version of the NML criterion [Kontkanen and Myllymäki, 2007, Roos et al., 2008a] which corresponds to a decomposable NML score, $k^{\mathsf{NML}}_{\mathcal{G},\mathcal{D}} = \sum_{X_i} k^{\mathsf{NML}}_{X_i|\{\mathrm{Pa}_{X_i}\}}$, defined as,

$$k^{\mathsf{NML}}_{Y|\{\mathrm{Pa}_Y\}} = \sum_{j}^{q_y} \log \mathcal{C}^{r_y}_{N_{yj}} \tag{6.23}$$

$$\Delta k^{\mathsf{NML}}_{Y|\{\mathrm{Pa}_Y\}\setminus X} = \sum_{j}^{q_y} \log \mathcal{C}^{r_y}_{N_{yj}} - \sum_{j'}^{q_y/r_x} \log \mathcal{C}^{r_y}_{N_{yj'}} \tag{6.24}$$

where N_{yj} is the number of data points corresponding to the jth state of the parents of Y, $\{\mathrm{Pa}_Y\}$, and $N_{yj'}$ the number of data points corresponding to the j'th state of the parents of Y, excluding X, $\{\mathrm{Pa}_Y\}_{\setminus X}$. Hence, the factorized NML score for each node X_i corresponds to a separate normalisation for each state $j = 1, ..., q_i$

of its parents and involving exactly N_{ij} data points of the finite dataset,

$$\begin{aligned}\mathcal{L}_{\mathcal{G}} &= e^{-N\sum_i H(X_i|\{\mathrm{Pa}_{X_i}\})-\sum_i\sum_j^{q_i}\log\mathcal{C}_{N_{ij}}^{r_i}} \\ &= e^{N\sum_i\sum_j^{q_i}\sum_k^{r_i}\frac{N_{ijk}}{N}\log\left(\frac{N_{ijk}}{N_{ij}}\right)-\sum_i\sum_j^{q_i}\log\mathcal{C}_{N_{ij}}^{r_i}} \\ &= \prod_i\prod_j^{q_i}\frac{\prod_k^{r_i}\left(\frac{N_{ijk}}{N_{ij}}\right)^{N_{ijk}}}{\mathcal{C}_{N_{ij}}^{r_i}}\end{aligned}$$

where N_{ijk} corresponds to the number of data points for which the ith node is in its kth state and its parents in their jth state, with $N_{ij}=\sum_k^{r_i}N_{ijk}$. The universal normalization constant $\mathcal{C}_n^r$ is then obtained by averaging over all possible partitions of the n data points into a maximum of r subsets, $\ell_1+\ell_2+\cdots+\ell_r=n$ with $\ell_k\geqslant 0$,

$$\mathcal{C}_n^r=\sum_{\ell_1+\ell_2+\cdots+\ell_r=n}\frac{n!}{\ell_1!\ell_2!\cdots\ell_r!}\prod_{k=1}^{r}\left(\frac{\ell_k}{n}\right)^{\ell_k} \tag{6.25}$$

which can in fact be computed in linear-time using the following recursion [Kontkanen and Myllymäki, 2007],

$$\mathcal{C}_n^r=\mathcal{C}_n^{r-1}+\frac{n}{r-2}\mathcal{C}_n^{r-2} \tag{6.26}$$

with $\mathcal{C}_0^r=1$ for all r, $\mathcal{C}_n^1=1$ for all n and applying the general formula Eq.6.25 for $r=2$,

$$\mathcal{C}_n^2=\sum_{h=0}^{n}\binom{n}{h}\left(\frac{h}{n}\right)^h\left(\frac{n-h}{n}\right)^{n-h} \tag{6.27}$$

or its Szpankowski approximation for large n (needed for $n>1000$ in practice) [Kontkanen, 2009, Kontkanen et al., 2003, Szpankowski, 2001],

$$\mathcal{C}_n^2=\sqrt{\frac{n\pi}{2}}\left(1+\frac{2}{3}\sqrt{\frac{2}{n\pi}}+\frac{1}{12n}+\mathcal{O}\left(\frac{1}{n^{3/2}}\right)\right) \tag{6.28}$$

$$\simeq\sqrt{\frac{n\pi}{2}}\exp\left(\sqrt{\frac{8}{9n\pi}}+\frac{3\pi-16}{36n\pi}\right) \tag{6.29}$$

6.2.3 Probability estimate of indirect contributions

The previous results enable us to estimate the probability of a node Z to contribute to the conditional mutual information $I(X;Y|\{U_i\})$, by combining the probability,

$P_{\mathsf{nv}}(XYZ|\{U_i\})$, that the triple XYZ is a generalized non-v-structure conditioned on $\{U_i\}$ and the probability, $P_{\mathsf{b}}(XY|\{U_i\})$, that its base is XY, where,

$$P_{\mathsf{nv}}(XYZ|\{U_i\}) = \frac{\mathcal{L}_{\mathsf{nv}}(XY)}{\mathcal{L}_{\mathsf{nv}}(XY) + \mathcal{L}_{\mathsf{v}}(XY)}$$
$$P_{\mathsf{b}}(XY|\{U_i\}) = \frac{\mathcal{L}_{\mathsf{nv}}(XY)}{\mathcal{L}_{\mathsf{nv}}(XY) + \mathcal{L}_{\mathsf{nv}}(XZ) + \mathcal{L}_{\mathsf{nv}}(YZ)}$$

that is, using Eqs. 6.4 & 6.5 and including finite size corrections of the maximum likelihoods (*i.e.* shifted information from Eqs. 6.16 & 6.15),

$$P_{\mathsf{nv}}(XYZ|\{U_i\}) = \frac{1}{1 + e^{-NI'(X;Y;Z|\{U_i\})}} \tag{6.30}$$
$$P_{\mathsf{b}}(XY|\{U_i\}) = \frac{1}{1 + \frac{e^{-NI'(X;Z|\{U_i\})}}{e^{-NI'(X;Y|\{U_i\})}} + \frac{e^{-NI'(Y;Z|\{U_i\})}}{e^{-NI'(X;Y|\{U_i\})}}} \tag{6.31}$$

Then, various alternatives to combine $P_{\mathsf{nv}}(XYZ|\{U_i\})$ and $P_{\mathsf{b}}(XY|\{U_i\})$ exist to estimate the overall probability that the additional node Z indirectly contributes to $I(X;Y|\{U_i\})$. One possibility is to choose the lower bound $S_{\mathsf{lb}}(Z;XY|\{U_i\})$ of $P_{\mathsf{nv}}(XYZ|\{U_i\})$ and $P_{\mathsf{b}}(XY|\{U_i\})$ as a score, since both conditions need to be fulfilled to warrant that Z indeed contributes to $I(X;Y|\{U_i\})$,

$$S_{\mathsf{lb}}(Z;XY|\{U_i\}) = \min\left[P_{\mathsf{nv}}(XYZ|\{U_i\}), P_{\mathsf{b}}(XY|\{U_i\})\right]$$

The pair of nodes XY with the most likely contribution from a third node Z can then be ordered according to their rank $R(XY;Z|\{U_i\})$ defined as,

$$R(XY;Z|\{U_i\}) = \max_{Z}\left(P_{\mathsf{lb}}(Z;XY|\{U_i\})\right) \tag{6.32}$$

and Z can be iteratively added to the set of contributing nodes (*i.e.* $\{U_i\} \leftarrow \{U_i\} + Z$) of the top link $XY = \text{argmax}_{XY} R(XY;Z|\{U_i\})$ to progressively recover the most significant indirect contributions to all pairwise mutual information in a causal graph, as outlined below.

6.3 The 3off2 scheme

6.3.1 Robust inference of conditional independencies

The previous results can be used to provide a robust inference method to identify conditional independencies and, hence, reconstruct the skeleton of underlying causal graphs from finite available observational data. The approach follows the spirit of constraint-based methods, such as the PC or IC algorithms, but recovers conditional independencies following an evolving ranking of the network edges, $R(XY;Z|\{U_i\})$, defined in Eq. 6.32.

Then, Z can be iteratively added to the set of contributing nodes (*i.e.* $\{U_i\} \leftarrow \{U_i\}+Z$) of the top edge $XY = \text{argmax}_{XY} R(XY;Z|\{U_i\})$ to progressively recover the most significant indirect contributions to all pairwise mutual information in a causal graph.

Implementing this local optimization scheme, the 3off2 algorithm eventually learns the network skeleton by collecting the nodes of the separation sets one-by-one, instead of exploring the full combinatorics of sepset composition without any likelihood guidance. Indeed, the 3off2 scheme amounts to identify $\{U_i\}$ by "taking off" iteratively the "most likely" conditional 3-point information from each 2-point information as,

$$\begin{aligned} I(X;Y|\{U_i\}_n) = I(X;Y) &- I(X;Y;U_1) \\ &- I(X;Y;U_2|U_1) - \cdots \\ &- I(X;Y;U_n|\{U_i\}_{n-1}) \end{aligned}$$

or equivalently between the shifted 2-point and 3-point information terms, defined in Eqs. 6.15 and 6.16 respectively,

$$\begin{aligned} I'(X;Y|\{U_i\}_n) = I'(X;Y) &- I'(X;Y;U_1) \\ &- I'(X;Y;U_2|U_1) - \cdots \\ &- I'(X;Y;U_n|\{U_i\}_{n-1}) \end{aligned} \tag{6.33}$$

This leads to the Algorithm 4 for the reconstruction of the graph skeleton using the 3off2 scheme. Note, in particular, that the 3off2 scheme to reconstruct graph skeleton is solely based on identifying structural independencies, which can also be applied to graphical models for undirected Markov networks.

Algorithm 4: 3off2 Network Reconstruction

In: finite observational dataset of size N;
complexity $k_{X;Y|\{U_i\}}$
Out: (partially) oriented graph $\mathcal{G}$
0. Initiation
Start with complete undirected graph $\mathcal{G}$
forall the *links* XY **do**
  if $I(X;Y) < k_{X;Y|\emptyset}/N$ i.e. $I'(X;Y) < 0$ **then**
    XY **link is** non-essential and **removed**
    separation set of XY: $\mathrm{Sep}_{XY} = \emptyset$
  else
    find the **most contributing node** Z neighbor of X or Y and **compute** 3off2 **rank**, $R(XY;Z|\emptyset)$
  end
end
1. Iteration
while $\exists$ XY *link with* $R(XY;Z|\{U_i\}) > 1/2$ **do**
  for *top link* XY *with highest rank* $R(XY;Z|\{U_i\})$ **do**
    expand contributing set $\{U_i\} \leftarrow \{U_i\} + Z$
    if $I(X;Y|\{U_i\}) < k_{X;Y|\{U_i\}}/N$ i.e. $I'(X;Y|\{U_i\}) < 0$ **then**
      XY **link is** non-essential and **removed**
      separation set of XY: $\mathrm{Sep}_{XY} = \{U_i\}$
    else
      find **next most contributing node** Z neighbor of X or Y and **compute new** 3off2 **rank** of XY: $R(XY;Z|\{U_i\})$
    end
    sort the 3off2 **rank list** $R(XY;Z|\{U_i\})$
  end
end
2. Orientation / Propagation
Sort list of unshielded triples, $\mathcal{L}_c = \{\langle X,Z,Y\rangle_{X\neq Y}\}$,
in decreasing order of $|I'(X;Y;Z|\{U_i\})|$
repeat
  Take $\langle X,Z,Y\rangle_{X\neq Y} \in \mathcal{L}_c$ with highest $|I'(X;Y;Z|\{U_i\})|$ on which R_0 or R_1 orientation rule can be applied
  if $I'X;Y;Z|\{U_i\}) < 0$ **then**
    if $\langle X,Z,Y\rangle_{X\neq Y}$ has no diverging orientation, apply
    R_0: $\{X - * Z * - Y \;\&\; X \neq Y \;\&\; Z \notin \mathrm{Sep}_{XY}\} \Rightarrow \{X \rightarrow Z \leftarrow Y\}$
  else
    if $\langle X,Z,Y\rangle_{X\neq Y}$ has one converging orientation, apply
    R_1: $\{X \rightarrow Z - Y \;\&\; X \neq Y\} \Rightarrow \{Z \rightarrow Y\}$
  end
  Apply new orientation(s) to all other $\langle X',Z',Y'\rangle_{X'\neq Y'} \in \mathcal{L}_c$
until *no additional orientation can be obtained*;

6.3.2 The 3off2 algorithm

The 3off2 scheme can be used to devise a two-step algorithm (see Algorithm (4)), inspired by constraint-based approaches, to first reconstruct network skeleton (Algorithm (4), step 1) before combining orientation and propagation of edges in a single step based on likelihood ratios (Algorithm (4), step 2).

6.3.2.1 Reconstruction of network skeleton

The 3off2 scheme will first be applied to iteratively remove edges with *maximum positive contributions*, $I'(X;Y;U_k|\{U_i\}_{k-1}) > 0$, corresponding to the *most likely generalized non-v-structures* (Eq. 6.4), while *minimizing simultaneously* the remaining 2-point information, $I'(X;Y|\{U_i\}_k)$ (Eq. 6.5), consistently with the data processing inequality.

Such 3off2 scheme (Algorithm (4), step 1) will therefore progressively lower the conditional 2-point information terms, $I'(X;Y) > \cdots > I'(X;Y|\{U_i\}_{k-1}) > I'(X;Y|\{U_i\}_k)$ and might ultimately result in the removal of the corresponding edge, XY, but only when a structural independency is actually found, *i.e.* $I'(X;Y|\{U_i\}_n) < 0$, as in constraint-based algorithms for a given significance level α. Yet, the skeleton obtained with the 3off2 scoring approach is expected to be more robust to finite observational data than the skeleton obtained with PC or IC algorithms, as the former results only from statistically significant 3-point contributions, $I'(X;Y;U_k|\{U_i\}_{k-1}) > 0$, based on their quantitative 3off2 ranks, $R(XY;U_k|\{U_i\}_{k-1})$.

The best results on benchmark networks using these quantitative 3off2 ranks are obtained with the NML score. The MDL score leads to equivalent results, as expected, in the limit of very large datasets (Chapter 7). However, with smaller datasets, the most reliable results with the MDL score are obtained using *non-shifted* instead of shifted 2-point and 3-point information terms in the 3off2 rank of individual edges, Eq.6.32. This is because the MDL complexity tends to underestimate the importance of edges between nodes with many levels (section 6.2.2). For finite datasets, it easily leads to spurious conditional independencies, $I'(X;Y|Ui) < 0$, when using shifted 2-point and 3-point information, Eq.6.33, whereas using non-shifted information in the 3off2 ranks (Eq.6.32) tends to limit the number of false negatives as early errors in $\{U_i\}$ can only increase $I(X;Y|Ui) \geqslant 0$, in the end.

6.3.2.2 Orientation of network skeleton

The skeleton and the separation sets resulting from the 3off2 iteration step (Algorithm (4), step 1) can then be used to orient the edges and to propagate orientations to the unshielded triples. However, while the constraint-based methods distinguish the v-structures orientation step (Algorithm (1), step 2) from the propagation procedure (Algorithm 1, step 3), the 3off2 algorithm intertwines these two steps based on the respective likelihood scores of individual v-structures and non-v-structures (Algorithm (4), step 2).

As stated earlier, the magnitude and sign of the conditional 3-point information, $I(X;Y;Z|\{U_i\})$ (or equivalently the shifted 3-point information, Eq. 6.4), indicate if a non v-structure is more likely than a v-structure. Hence, all the unshielded triples can be ranked by the *absolute* value of their conditional 3-point information, that is, in decreasing order of their *likelihood* of being either a v-structure or a non-v-structure.

As detailed in the step 2 of Algorithm (4), the most likely v-structure is used to set the first orientations, following R_0 orientation rule. The possible propagations are then performed, following R_1 propagation rule, starting from the unshielded triple having the most positive conditional 3-point information. The following most likely v-structure is considered when no further propagation is possible on unshielded triples with greater absolute 3-point information. If conflicting orientations arise (such as $a \rightarrow b \leftarrow c$ & $b \rightarrow c \leftarrow d$), the less likely v-structure and its possible propagations are ignored, unless latent variables are considered (section 6.4.2).

Note that we only implement the R_0 and R_1 propagation rules, which are applied in decreasing order of likelihood. In particular, we do not consider propagation rules R_2 and R_3 which are not associated to likelihood scores but enforce the hypothesis of acyclic constraint.

As for the 3off2 skeleton reconstruction, the orientation/propagation step of 3off2 allows for a robust discovery of orientations from finite observational data as it relies on statistically significant conditional 3-point information, $I'(X;Y;Z|\{U_i\})$. During this step, 3off2 recovers and propagates as many orientations as possible in an iterative procedure following the quantitative ranks of the unshielded triples based on their conditional 3-point information.

6.4 Extension of the 3off2 algorithm

6.4.1 Probabilistic estimation of the edge endpoints

Given the skeleton obtained from Algorithm 4, Eqs.6.30 and 6.31 lead to the following Proposition 7 and Proposition 8 for the orientation and propagation rules of unshielded triples, which are equivalent to the Propostion 5 and Proposition 6 but for underlying DAG models (assuming no latent variables) and for finite datasets with the corresponding probabilities for the initiation/propagation of orientations.

Proposition 7 (Significantly negative conditional 3-point information as robust statistical evidence of causality in finite datasets). *Assuming that the underlying graphical model is a DAG $\mathcal{G}$ on V, $\forall X, Y, Z \in V$ and $\forall \{U_i\} \subseteq V$ s.t. $I'(X;Y|\{U_i\}) < 0$ (i.e. no XY edge) and $I'(X;Y;Z|\{U_i\}) < 0$ then,*

i. if X, Y, Z form an unshielded triple, $X \circ\!\!-\!\!\circ Z \circ\!\!-\!\!\circ Y$, then it should be oriented as $X \to Z \leftarrow Y$, with probabilities,

$$P^{\circ}_{X\to Z} = P^{\circ}_{Y\to Z} = \frac{1 + e^{NI'(X;Y;Z|\{U_i\})}}{1 + 3e^{NI'(X;Y;Z|\{U_i\})}}$$

ii. similarly, if X, Y, Z form an unshielded triple, with one already known converging arrow, $X \to Z \circ\!\!-\!\!\circ Y$, with probability $P_{X\to Z} > P^{\circ}_{X\to Z}$, then the second edge should be oriented to form a v-structure, $X \to Z \leftarrow Y$, with probability,

$$P_{Y\to Z} = P_{X\to Z}\left(\frac{1}{1 + e^{NI'(X;Y;Z|\{U_i\})}} - \frac{1}{2}\right) + \frac{1}{2}$$

Proof. The implications (*i.*) and (*ii.*) rely on Eq.6.30 to estimate the probability that the two edges form a collider ($P_{\mathsf{v}}(XYZ|\{U_i\}) = 1 - P_{\mathsf{nv}}(XYZ|\{U_i\})$). We start proving (*ii.*) using the probability decomposition formula:

$$\begin{aligned} P_{Y\to Z} &= P_{X\to Z}\frac{P_{X\to Z\leftarrow Y}}{P_{X\to Z\leftarrow Y} + P_{X\to Z\to Y}} \\ &\quad + (1 - P_{X\to Z})\frac{P_{X\leftarrow Z\leftarrow Y}}{P_{X\leftarrow Z\leftarrow Y} + P_{X\leftarrow Z\to Y}} \\ &= P_{X\to Z}\left(\frac{1}{1 + e^{NI'(X;Y;Z|\{U_i\})}} - \frac{1}{2}\right) + \frac{1}{2} \end{aligned}$$

which also leads to (*i.*) if one assumes $P_{X\to Z} = P_{Y\to Z}$ by symmetry in absence of prior information on these orientations. □

Following the rationale of constraint-based approaches, it is then possible to 'propagate' further the orientations downstream of colliders, using Eq.6.30 for positive (conditional) 3-point information. For simplicity and consistency, we only implement the propagation of orientation based on likelihood ratios, which can be quantified for finite datasets as proposed in the following Proposition 8. In particular, we do not extend the propagation rules [Meek, 1995] to inforce acyclic constraints that are necessary to have a complete reconstruction of the Markov equivalent class of the underlying DAG model.

Proposition 8 (robust 'propagation' of causality at unshielded triples with significantly positive conditional 3-pt information).
Assuming that the underlying graphical model is a DAG $\mathcal{G}$ *on* V, $\forall X, Y, Z \in V$ *and* $\forall \{U_i\} \subseteq V_{\setminus\{X,Y,Z\}}$ *s.t.* $I'(X;Y|\{U_i\},Z) < 0$ *(i.e. no* XY *edge) and* $I'(X;Y;Z|\{U_i\}) > 0$*, then if* X, Y, Z *form an unshielded triple with one already known converging orientation,* $X \to Z \circ\!\!-\!\!* Y$*, with probability* $P_{X\to Z} > 1/2$*, this orientation should be 'propagated' to the second edge as* $X \to Z \to Y$*, with probability,*

$$P_{Z\to Y} = P_{X\to Z}\left(\frac{1}{1+e^{-NI'(X;Y;Z|\{U_i\})}} - \frac{1}{2}\right) + \frac{1}{2}$$

Proof. This results is shown using the probability decomposition formula,

$$\begin{aligned} P_{Z\to Y} &= P_{X\to Z}\frac{P_{X\to Z\to Y}}{P_{X\to Z\leftarrow Y} + P_{X\to Z\to Y}} \\ &\quad + (1 - P_{X\to Z})\frac{P_{X\leftarrow Z\to Y}}{P_{X\leftarrow Z\leftarrow Y} + P_{X\leftarrow Z\to Y}} \\ &= P_{X\to Z}\left(\frac{1}{1+e^{-NI'(X;Y;Z|\{U_i\})}} - \frac{1}{2}\right) + \frac{1}{2} \end{aligned}$$

□

Proposition 7 and Proposition 8 lead to the following Algorithm 5 for the orientation of unshielded triples of the graph skeleton obtained from Algorithm 4.

Algorithm 5: 3off2 probabilistic Orientation / Propagation Step

In: Graph skeleton from Algorithm 4 (step 1) and corresponding conditional 3-point information $I'(X;Y;Z|\{U_i\})$.

Out: Partially oriented causal graph $\mathcal{G}$ with edge orientation probabilities.

3off2 probabilistic Orientation / Propagation Step

sort list of unshielded triples, $\mathcal{L}_c = \{\langle X,Z,Y\rangle_{X\neq Y}\}$, in decreasing order of their orientation/propagation probability initialized at 1/2 and computed from:
- (*i.*) Proposition 7, if $I'(X;Y;Z|\{U_i\})<0$, or
- (*ii.*) Proposition 8, if $I'(X;Y;Z|\{U_i\})>0$

repeat

  Take $\langle X,Z,Y\rangle_{X\neq Y} \in \mathcal{L}_c$ with highest orientation / propagation probability $> 1/2$.

  if $I'(X;Y;Z|\{U_i\}) < 0$ **then**

    Orient/propagate edge direction(s) to form a **v-structure** $X \to Z \leftarrow Y$ with probabilities $P_{X\to Z}$ and $P_{Y\to Z}$ given by **Proposition 7**.

  else

    Propagate second edge direction to form a **non-v-structure** $X \to Z \to Y$ assigning probability $P_{Z\to Y}$ from **Proposition 8**.

  end

  Apply new orientation(s) and **sort** remaining list of unshielded triples $\mathcal{L}_c \leftarrow \mathcal{L}_c \backslash \langle X,Z,Y\rangle_{X\neq Y}$ after **updating propagation probabilities**.

until *no additional orient./propa. probability* $>1/2$;

6.4.2 Allowing for latent variables

Although constraint-based methods have benefited from the recent improvements of the PC algorithm (see PC-stable version in section 3.2.1), learning causal graphical models from observational data generated by complex biological systems frequently requires to allow for the presence of unobserved latent variables. As introduced in section 3.2.2, several constraint-based methods, such as the FCI ('Fast Causal Inference') [Spirtes et al., 1999, 2000], the RFCI ('Really Fast Causal Inference') [Colombo et al., 2012] or the AnytimeFCI [Spirtes, 2001] algorithms can learn a PAG (Partial Ancestral Graph, see definition 2.2.5) over the observed variables while taking into account the presence of latent variables. As exemplified in Fig. 3.1 section 3.2.2.1, when the causal sufficiency assumption does not hold, one should also consider nodes that do not belong to the neighbourhood of X or Y as potential members of the separation set, Sep_{XY}, of this two variables. Indeed,

in the example given in Fig. 3.1, the presence of the two latent variables, L_1 and L_2, implies that X_3 should be added to the separation set of the pair (X_1, X_5) in order to enable the removal of the edge $X_1 - X_5$, as done in the PAG of the true underlying causal network.

By contrast, the 3off2 algorithm enables a simple extension of the search of the possible contributors to a pair of variables (X, Y) including variables that do not belong to the neighbourhood of X or Y. This is done through a simple modification of lines 11 and 22 in the Algorithm 4, by setting the 3off2 approach to allow the search of $\mathbf{Z}$ beyond the neighbourhood of X or Y. This modification implies to check a greater number of nodes when progressively building the separation set. Yet, as no combinatorial search of contributing nodes is done in the 3off2 approach, the iterative search of the most likely contributing node can be done beyond the first neighbours without significant computational burden (and thus avoiding the multiple steps of the FCI or related algorithms and the use of 'inducing paths' to uncover likely non neighbours contributing nodes).

As shown in Chapter 8, the 3off2 approach could be used to reconstruct networks of genomic properties suspected to have played a role in the evolution of the human genome and its enhanced susceptibility to genetic disease and cancer. In particular, the networks learnt with the 3off2 method support our non-adaptive evolutionary model which posits that the retention of copies of genes arising from whole-genome duplication events is primarily the consequence of their susceptibility to dominant deleterious mutations [Singh et al., 2012]. However, many other properties may have participated in the fate of duplicated genes, and some of them are still unknown. For instance, Fig. 8.4 exhibits a bi-directed edge between the nodes labelled 'Essential' and 'Ka.Ks', which suggests that there remains hidden latent cause(s) beyond those included in the model (disease, SSD...) to relate a gene *essentiality* to its tendency to be (strongly) conserved. Another example is given in Chapter 9, where the 3off2 method is used to reconstruct networks corresponding to 2D layers of a zebrafish brain, corresponding to different depths of the brain. As neural regions usually span over the three dimensions, using the 3off2 method while allowing for the presence of latent variables, *i.e.* neurons belonging to the upper or lower layers, can prevent us from drawing spurious cause-effect relationships within single layers. Finally, Chapter 11, dedicated to the reconstruction of mutational pathways in breast cancer, also shows the importance of assuming the presence of latent variables. Indeed, the restricted number of samples does not allow to take into account thousands of genes in the network reconstructions. Thus, based on specific biological parameters, one has to define a set of

genes suspected to participate in the tumour progression, which leads to ignore genes that may play a role in the emergence and proliferation of malignant cells. In particular, as shown in Fig. 11.1, the relationship between the genes CDH1 and ERBB2 may not be a cause-effect interaction as suspected in recent studies, but the consequence of hidden common cause(s).

6.4.3 Allowing for missing data

Researchers need frequently to deal with missing values when using data generated from *real-life* systems. Missing data can have diverse sources such as mistakes during the records or missing answer to a survey. Methods such as Bayesian search-and-score approaches which require a global score need to rely on complete information without missing data. When it is not the case, the missing data need to be somehow generated from the available data with obvious statistical issues. Filling the missing information could be done for instance using a simple mean value over the available data, or a more complex statistical model. However, if there is no bias in the missing data, *i.e.* if the data are *missing completely at random* (MCAR, [Rubin, 1976]), one can assume that the available information is still representative of the population. In this case, the counterpart of relying only on the available information is that all samples will not contribute equally to the model with concomitant variations in statistical precision of different subnetworks reconstructed from the available data.

The 3off2 approach provides a way to rely only on the available information and estimate local confidence of the reconstructed network. As the 3off2 method only requires the calculation of $I(X;Y|\{U_i\})$ and $I(X;Y;Z|\{U_i\})$ at each iteration, the (sub)dataset considered needs to be complete solely for the variables $X, Y, Z, \{U_i\}$ involved in the calculation of these terms.

As exemplified in Chapters 8 and 11, this possibility to deal with missing data with the 3off2 method is shown to be crucial in the network reconstruction of genomic properties and mutational pathways. For the application detailed in Chapter 11, the gene network reconstructions rely on online databases such as COSMIC (*Catalogue Of Somatic Mutations In Cancer*, [Forbes et al., 2008]). Although the COSMIC database contains over a million of samples, only about 1% of these samples correspond to whole-exome sequencing. Indeed, most samples were obtained before the recent advent of NGS technologies. Besides, for a specific cancer type, the number of samples can drop to only a few tens. As illustrated in Table 6.1, if

one decide to rely solely on complete information, the only sample available would be $samp_4$ (*i.e.* 1% of the dataset). However, genes g_3 and g_7 have in fact been simultaneously sequenced in 3 samples ($samp_1$, $samp_3$ and $samp_4$), and genes g_2 and g_7 in 2 samples ($samp_1$ and $samp_4$).

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	
$samp_1$	n		y				n	
$samp_2$		y					y	
$samp_3$			n				n	
$samp_4$	n	n	y	n	n	y	n	$\leftarrow 1\%$
$samp_5$		n			n			

TABLE 6.1 **Example of a dataset with missing information.** The missing values can drastically reduce the number of samples if a *globally* complete dataset is required.

The 3off2 approach is taking the most out of incomplete datasets by allowing the computations to be performed only on the partially complete samples relative to the restricted set of variables considered (*i.e.* X, Y, Z and $\{U_i\}$).

6.5 Implemented pipelines

6.5.1 The 3off2 pipeline

The 3off2 reconstruction method has been implemented mainly in R, and partly in C. Fig. 6.2 gives an overview of the pipeline through which evaluations of the 3off2 approach have been performed. This pipeline takes as compulsory inputs (i) a dataset containing N samples of p variables (discretized data) and (ii) an output directory path. The complexity can follow an MDL criterion, rather than the default NML criterion (see section 6.2.2 for details on MDL and NML graph complexities). If the samples of the input dataset are correlated, and if the number of independent samples is known (effective sample size, N_{eff}), this information can also be provided to the pipeline (see section 7.3, related to the reconstruction of undirected networks, that gives further details and examples for this input argument). The parameter η is an integer (default value 1) that modifies the severity of the removal decision (higher values inducing more strongly biased decision towards removal). This amounts to shift the likelihood ratio Eq. 6.12 as,

$$\frac{\mathcal{L}_{\mathcal{G}_{\setminus XY|\{U_i\}}}}{\mathcal{L}_{\mathcal{G}}} = \eta e^{-NI(X;Y|\{U_i\})+k_{X;Y|\{U_i\}}} \tag{6.34}$$

which corresponds to redefining the two-point complexities as,

$$k'_{X;Y|\{U_i\}} = k_{X;Y|\{U_i\}} + \ln \eta$$

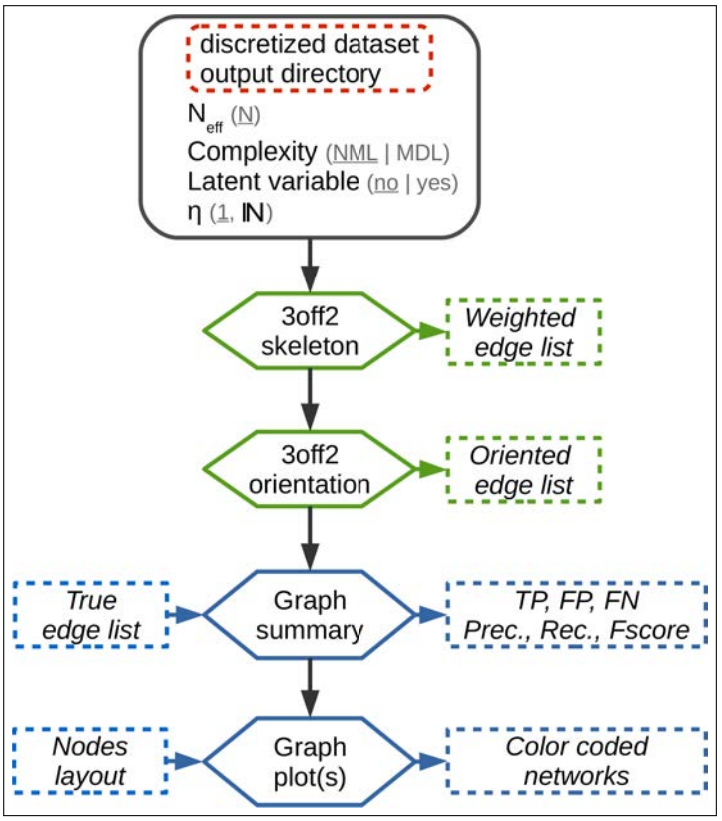


FIGURE 6.2 **The 3off2 reconstruction pipeline.**

As detailed in section 6.3.2, the 3off2 method first learns the skeleton of the graphical model ('3off2 skeleton' module), and then orients the edges ('3off2 orientation' module). These two steps provide the list of the discovered edges, as well as their conditional mutual information ($I(X;Y|\{U_i\})$), their set of contributors ($\{U_i\}$), the *local* complexity (associated with $\{X, Y, \{U_i\}\}$), the number of samples on which relies the conditional mutual information ($I(X;Y|\{U_i\})$), the orientation of the edge and the sign of the associated effect.

When the true underlying graphical model is known, it can be provided to the 'Graph summary' module that computes the number of true positive (TP), false positive (FP) and false negative (FN) edges (possibly adjusted with the number of edges wrongly oriented). The 'Graph summary' module also provides the associated Precision (Prec.), Recall (Rec.) and F-score (Fscore) measures. Finally, the 'Graph plot(s)' module, that can also take as input the nodes layout, generates several color coded plots. The edge color could either represent the *confidence* for

this edge (*i.e.* the difference between its conditional mutual information and the local complexity), or the strength and sign of the total effect associated with this relationship (see Chapters 8, 9 or 11 for color coded plot examples).

6.5.2 The discoNet pipeline

The 3off2 reconstruction method has been evaluated against alternative approaches, namely the PC-stable algorithm [Colombo and Maathuis, 2014, Spirtes and Glymour, 1991]), a Bayesian hill-climbing heuristics with random restarts [Chickering et al., 1995] and Aracne, an information-based reconstruction method [Margolin et al., 2006] (evaluation results are given in Chapter 7 and Appendices A & B). To facilitate batch evaluations, all these alternative methods and the 3off2 reconstruction approach have been encapsulated in the discoNet pipeline (Fig. 6.3).

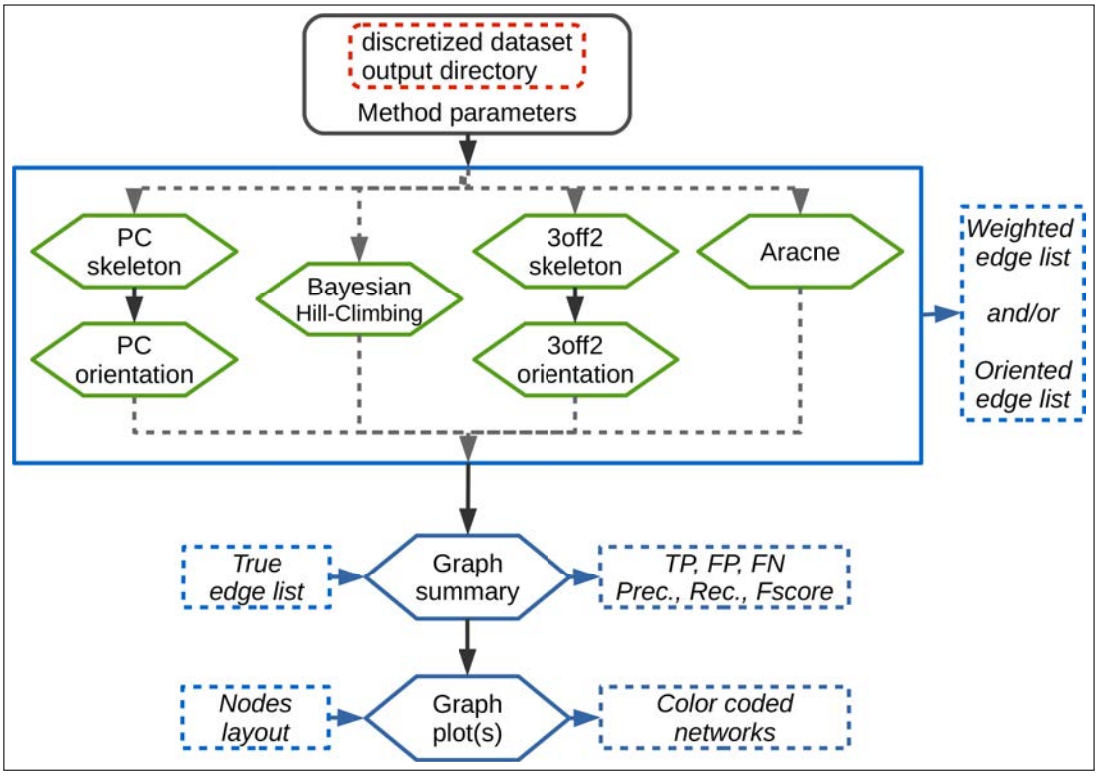


FIGURE 6.3 **The discoNet reconstruction pipeline.**

Parameters related to each method can be given as input arguments. For the PC algorithm, one can choose to run either the original [Spirtes and Glymour, 1991] or the stable version of the algorithm [Colombo and Maathuis, 2014]. The significance level α can also be provided, as well as the orientation rule (*i.e.* *conservative* [Ramsey et al., 2006] or *majority* [Colombo and Maathuis, 2014]). The implementation

of the PC algorithm corresponds to the `pcalg` R package [Kalisch and Bühlmann, 2008, Kalisch et al., 2012]. When using the hill-climbing approach, one can decide for the scoring function (BIC, Bayesian Information Criterion; BDe, Bayesian Dirichtlet equivalent; AIC, Akaike Information Criterion) and the number of random restarts. The implementation of the Bayesian hill-climbing heuristics in the discoNet pipeline corresponds to the `bnlearn` R package [Scutari, 2010]. Finally, the Aracne method can take into account a parameter ϵ to relax the removal of edges. The implementation of the Aracne method corresponds to the `minet` R package [Meyer et al., 2008]. The output files of these methods can then be given to the shared modules 'Graph summary' and 'Graph plot(s)', that provide comparable results.

Chapter 7

Evaluation on Benchmark Networks

This chapter gives the results of the 3off2 learning approach on different types of benchmarks networks. In section 7.1, the benchmark networks are *real-life* causal graphical models containing 20 to 70 nodes. The datasets have been simulated from known parameters. By contrast, in the section 7.2, the benchmark networks and their corresponding datasets have been generated using the causal modeling tool Tetrad IV.

The results are evaluated against other methods in terms of Precision (or positive predictive value), $Prec = TP/(TP + FP)$, Recall or Sensitivity (true positive rate), $Rec = TP/(TP + FN)$, as well as F-score $= 2 \times Prec \times Rec/(Prec + Rec)$ for increasing sample size N=10 to 50,000 data points. We also define additional Precision, Recall and F-scores taking into account the edge orientations of the predicted networks against the corresponding CPDAG of the benchmark networks. This amounts to label as false positives, all true positive edges of the skeleton with different orientation/non-orientation status as the CPDAG reference, $TP_{\text{misorient}}$, leading to the orientation-dependent definitions $TP' = TP - TP_{\text{misorient}}$ and $FP' = FP + TP_{\text{misorient}}$ with the corresponding CPDAG Precision, Recall and F-scores taking into account edge orientations.

7.1 Using simulated datasets from *real-life* networks

We have tested the 3off2 network reconstruction approach to learn benchmark causal graphs containing 20 to 70 nodes, Figures 7.1–7.5. The alternative inference methods used for comparison with 3off2 are the PC algorithm [Spirtes and Glymour, 1991] implemented in the `pcalg` package [Kalisch and Bühlmann, 2008, Kalisch et al., 2012] and Bayesian inference using the hill-climbing heuristics implemented in the `bnlearn` package [Scutari, 2010]. In addition, we also compare the skeleton of 3off2 to the unoriented output of Aracne [Margolin et al., 2006], an information-based inference approach, which iteratively prunes links with the weakest mutual information based on the Data Processing Inequality. We have used the Aracne implementation of the `minet` package [Meyer et al., 2008].

For each sample size, 3off2, Aracne, PC and the Bayesian inference methods have been tested on 50 replicates. Figures 7.1-7.5 give the average results over these multiple replicates when comparing the CPDAG (filled lines) of the reconstructed network (or its skeleton, dashed lined) to the CPDAG (or the skeleton) of the benchmark network.

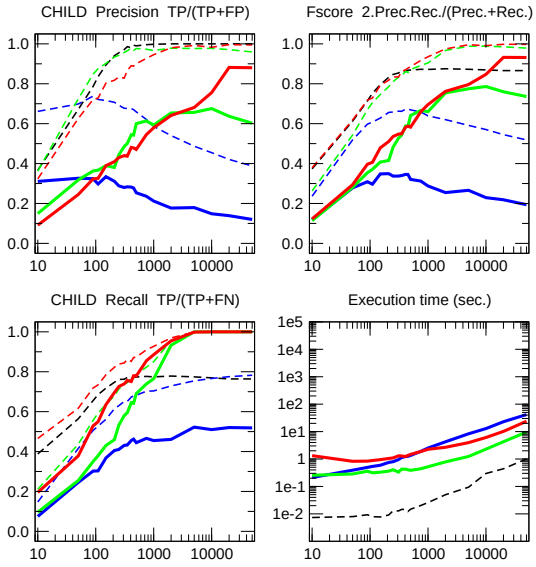


FIGURE 7.1 **CHILD network** [20 nodes, 25 links, 230 parameters, Average degree 2.5, Maximum in-degree 2]. Precision, Recall and F-score for skeletons (dashed lines) and CPDAGs (filled lines). The results are given for Aracne (black), PC (blue), Bayesian Hill-Climbing (green) and 3off2 (red).

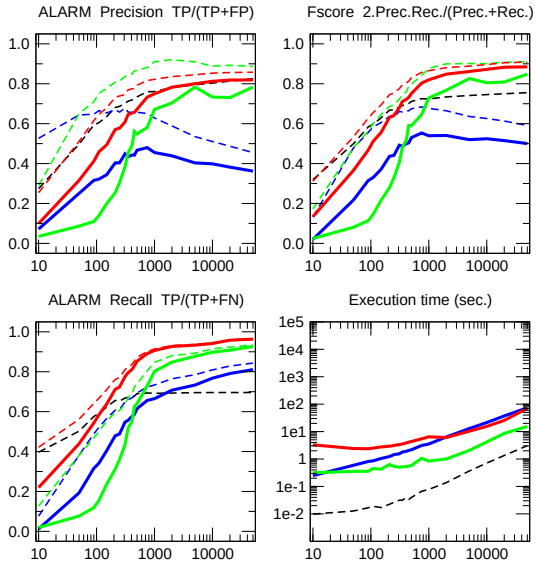


FIGURE 7.2 **ALARM network** [37 nodes, 46 links, 509 parameters, Average degree 2.49, Maximum in-degree 4]. Precision, Recall and F-score for skeletons (dashed lines) and CPDAGs (filled lines). The results are given for Aracne (black), PC (blue), Bayesian Hill-Climbing (green) and 3off2 (red).

For each method, the plots presented in Figures 7.1-7.5 are those obtained for the parameters that give overall the best results over the five reconstructed benchmark networks (see Figures A.1-A.20). In particular, we used the *stable* implementation of the PC algorithm, as well as the *majority rule* for the orientation and

propagation steps [Colombo and Maathuis, 2014]. PC's results are shown on Figures 7.1-7.5 for $\alpha = 0.1$. Decreasing α tends to improve the skeleton Precision at the expense of the skeleton Recall, leading in fact to worse skeleton F-scores for finite datasets, *e.g.* $N \leqslant 1000$ (see Figures A.1-A.5). The same trend is observed for CPDAG F-scores taking into account edge orientations, with best CPDAG scores at small sample sizes, obtained for larger α, *e.g.* $N \leqslant 1000$. Aracne threshold parameters for minimum difference in mutual information is set to $\epsilon = 0$, as small positive values typically worsen F-scores (see Figures A.6-A.10). Bayesian inference are obtained using BIC/MDL scores and hill-climbing heuristics with 100 random restarts [Chickering et al., 1995] (see Figures A.11-A.15).

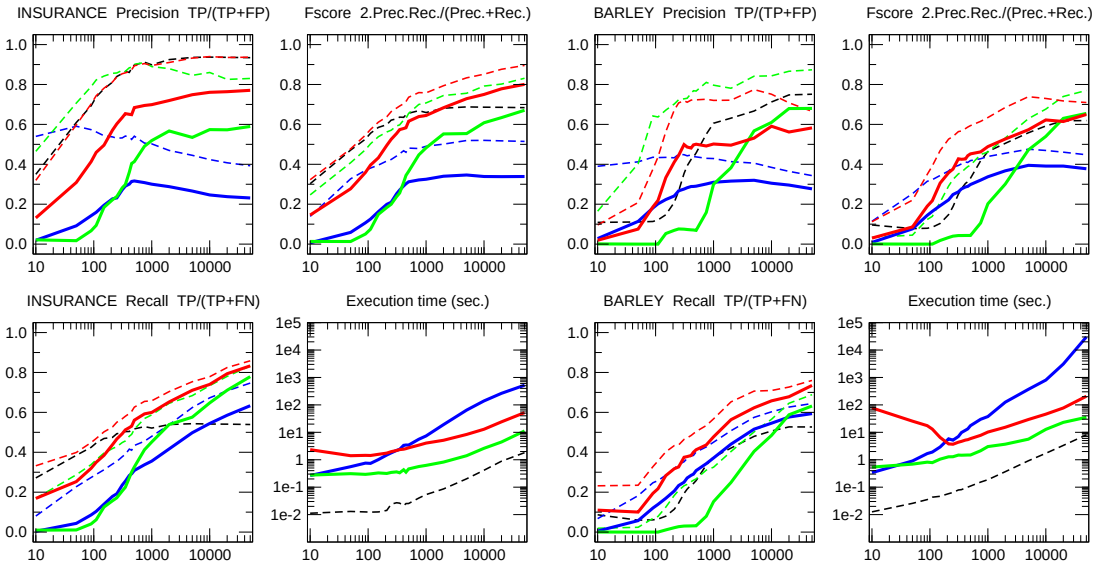


FIGURE 7.3 **INSURANCE network** [27 nodes, 52 links, 984 parameters, Average degree 3.85, Maximum in-degree 3]. Precision, Recall and F-score for skeletons (dashed lines) and CPDAGs (filled lines). The results are given for Aracne (black), PC (blue), Bayesian Hill-Climbing (green) and 3off2 (red).

FIGURE 7.4 **BARLEY network** [48 nodes, 84 links, 114,005 parameters, Average degree 3.5, Maximum in-degree 4]. Precision, Recall and F-score for skeletons (dashed lines) and CPDAGs (filled lines). The results are given for Aracne (black), PC (blue), Bayesian Hill-Climbing (green) and 3off2 (red).

Finally, the best 3off2 network reconstructions are obtained using NML scores with shifted 2-point and 3-point information terms in the rank of individual edges, see Methods. Using MDL scores, instead, leads to equivalent results, as expected, in the limit of very large datasets (see section 6.2.2). However, with smaller datasets, the most reliable results with MDL scores are obtained using *non-shifted* instead of shifted 2-point and 3-point information terms in the 3off2 rank of individual edges, as discussed in Methods (see Figures A.16-A.20).

All in all, we found that the 3off2 inference approach typically reaches better or equivalent F-scores for all dataset sizes as compared to all other tested methods,

i.e. Aracne, PC and Bayesian inference. This is clearly observed both on the skeletons (Figures 7.1 - 7.4 dashed lines) and even more clearly when taking the predictions of orientations into account (Figures 7.1-7.4 filled lines). Only the CPDAG F-score of the benchmark network HEPAR II (Figure 7.5), is slightly lower with 3off2 than with Bayesian inference at small sample sizes, although 3off2 eventually becomes slightly better for large datasets ($N \geqslant 4000$).

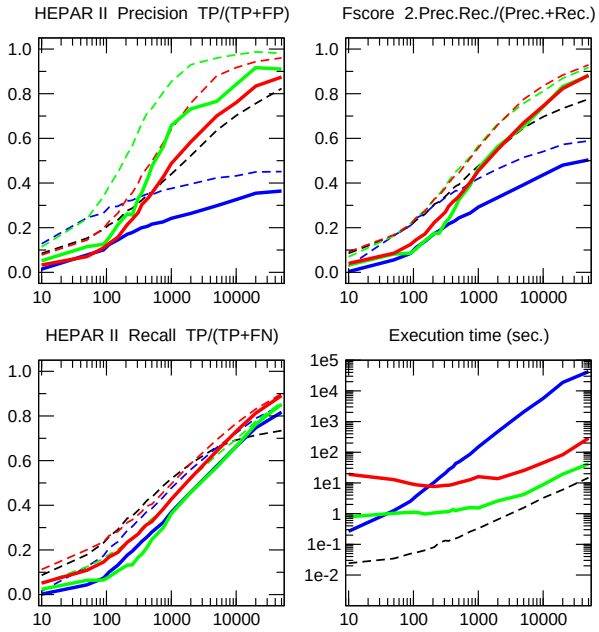


FIGURE 7.5 **HEPAR II network** [70 nodes, 123 links, 1,453 parameters, Average degree 3.51, Maximum in-degree 6]. Precision, Recall and F-score for skeletons (dashed lines) and CPDAGs (filled lines). The results are given for Aracne (black), PC (blue), Bayesian Hill-Climbing (green) and 3off2 (red).

7.2 Using simulated datasets from simulated networks

We have tested the 3off2 method on a range of benchmark networks of 50 nodes with up to 160 edges generated with the causal modeling tool Tetrad IV[1]. The average connectivity $\langle k \rangle$ of these benchmark networks ranges between 1.6 to 6.4, and the average maximal in/out-degree between 3.2 to 8.8 (see Table B.1 for a detailed description).

The alternative methods used for comparison with 3off2 are the PC algorithm implemented in the `pcalg` package [Kalisch and Bühlmann, 2008, Kalisch et al., 2012] and the hybrid method MMHC combining contraint-based skeleton and Bayesian orientation [Tsamardinos et al., 2006], implemented in the `bnlearn` package [Scutari, 2010]. Figures 7.6-7.10 give the average CPDAG comparison results over 100 dataset replicates from 5 different benchmark networks (Table S1). The causal graphical models predicted by the 3off2 method are obtained using either the MDL/BIC or the NML complexities (see section 6.2.2). Figures B.1-B.6 provide additional results on the prediction of the network skeletons and execution times.

The alternative predictions are based on the stable implementation of the PC algorithm as well as the majority and conservative rules for the orientation and propagation steps [Colombo and Maathuis, 2014]. The PC and MMHC results are shown, Figures 7.6-7.10, for an independence test parameter $\alpha = 0.1$, as reducing α tends to worsen the CPDAG F-score for benchmark networks with $\langle k \rangle \geq 1.6$[2] (Figures B.7-B.18). All in all, we found that the 3off2 method outperforms both PC-stable and MMHC methods on all tested datasets, Figures 7.6-7.10.

Additional comparisons were obtained with Bayesian inference implemented in the `bnlearn` package [Scutari, 2010], using AIC, BDe and BIC/MDL scores and hill-climbing heuristics with 30 to 100 random restarts, Figs B.19 - B.30. 3off2 reaches equivalent or significantly better F-scores than Bayesian hill-climbing for all dataset sizes on benchmark networks up to 120 edges ($\langle k \rangle = 4.8$). In particular, 3off2 with MDL scores reaches excellent F-scores on sparse networks (Figures B.19& B.20) and keeps one of the best F-scores over all sample sizes for less sparse networks when combined to NML complexity (Figures B.21-&-B.22). For somewhat denser networks ($\langle k \rangle \simeq 5$), the 3off2 F-score appears slightly lower than for Bayesian inference methods, Figure B.23, although it eventually becomes equivalent for large datasets ($N \geqslant 1000$).

[1] http://www.phil.cmu.edu/tetrad

[2] PC-stable benchmarks were tested up to $N = 10,000$ due to their sharp increase in execution time, see Figs. B.7-B.12.

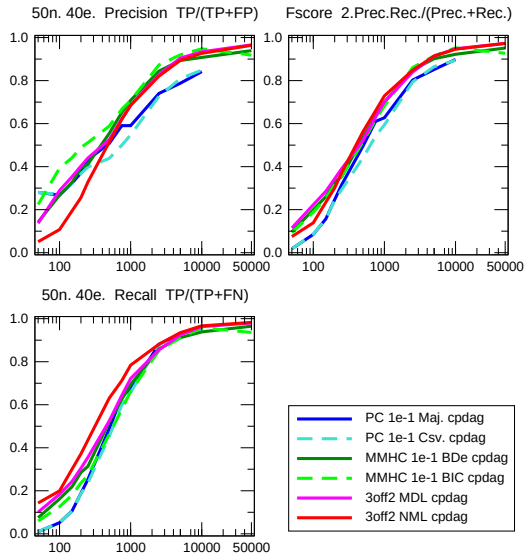


FIGURE 7.6 **50 node, 40 edge benchmark networks generated using Tetrad.** $\langle k \rangle = 1.6$, $\langle k^{in}_{\max} \rangle = 3.2$, $\langle k^{out}_{\max} \rangle = 3.6$.

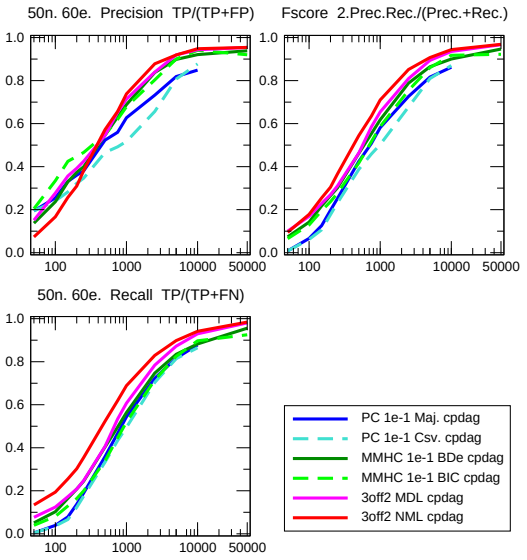


FIGURE 7.7 **50 node, 60 edge benchmark networks generated using Tetrad.** $\langle k \rangle = 2.4$, $\langle k^{in}_{\max} \rangle = 4.6$, $\langle k^{out}_{\max} \rangle = 3.6$.

On denser networks ($\langle k \rangle \geq 5-6$), Bayesian inference exhibits better F-scores than 3off2, in particular with AIC score, Figure B.24. However, the good performance with AIC strongly relies on its high Recall (but low Precision), due to its very small penalty term on large datasets, which makes it favor more complex networks (Figures B.24) but perform very poorly on sparse graphs (Figures B.19-B.21). By contrast, the reconstruction of dense networks is impeded by the 3off2 scheme, as it is not always possible to uncover structural independencies, $I(X;Y|\{U_i\}_n) \simeq 0$, in dense graphs through an ordered set $\{U_i\}_n$ with only positive conditional 3-point information, $I'(X;Y;U_k|\{U_i\}_{k-1}) > 0$.

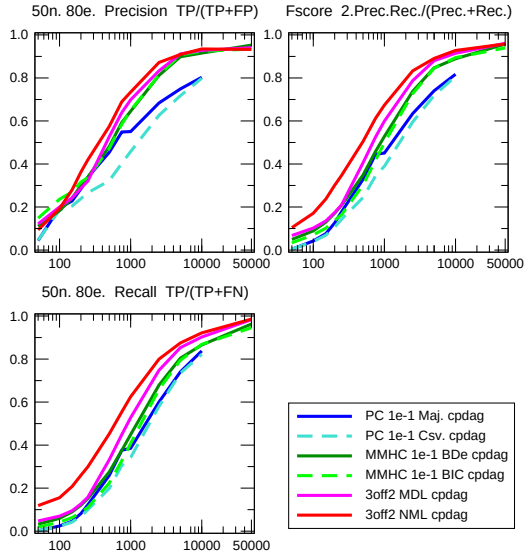


FIGURE 7.8 **50 node, 80 edge benchmark networks generated using Tetrad.** $\langle k \rangle = 3.2$, $\langle k^{in}_{\max} \rangle = 4.8$, $\langle k^{out}_{\max} \rangle = 5.6$.

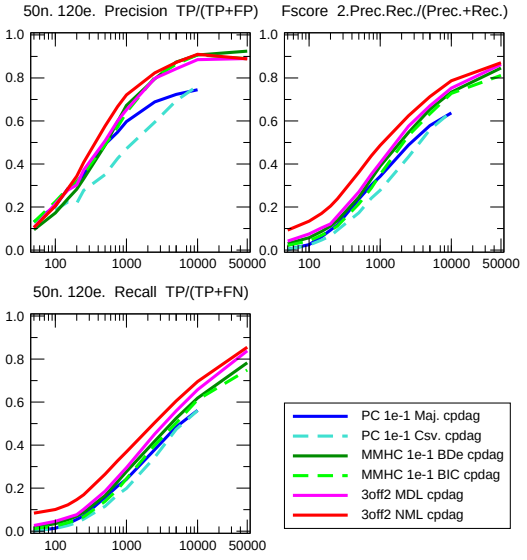


FIGURE 7.9 **50 node, 120 edge benchmark networks generated using Tetrad.** $\langle k \rangle = 4.8$, $\langle k^{in}_{\max} \rangle = 8.8$, $\langle k^{out}_{\max} \rangle = 7.2$.

Indeed in complex graphs, there are typically many indirect paths $X \to U_j \to Y$ between unconnected node pairs (X, Y). At the beginning of the pruning process, this is prone to suggest likely v-structures $X \to Y \leftarrow U_j$, instead of the correct non-v-structures, $X \to U_j \to Y$ (for instance if $I(X;U_j) \ll I(X;Y)$, $I(X;U_j) \ll I(U_j;Y)$ and $I(X;U_j) - I(X;U_j|Y) = I(X;Y;U_j) < 0$, for all j). Such elimination of FN edge $X \to U_j$ and conservation of FP $X \to Y$ tend to decrease both Precision and Recall, although 3off2 remains significantly more robust than PC and MMHC, Figure 7.10. Besides, for most practical applications on real life data, interpretable causal models should remain relatively sparse and avoid to display multiple indirected paths between unconnected nodes.

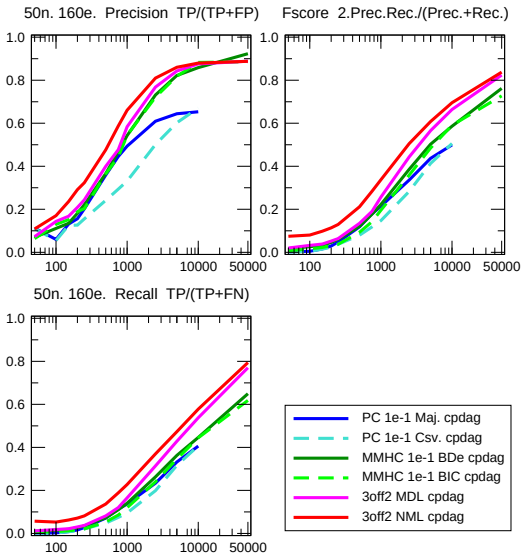


FIGURE 7.10 **50 node, 160 edge benchmark networks generated using Tetrad.** $\langle k \rangle = 6.4$, $\langle k^{in}_{\max} \rangle = 8.6$, $\langle k^{out}_{\max} \rangle = 8.6$.

Finally, 3off2 running times on these benchmark networks are similar to MMHC and Bayesian hill-climbing heuristic methods (with 100 restarts) and 10 to 100 times faster than PC for large datasets, Figures B.1- B.30.

7.3 Using simulated datasets from undirected networks

7.3.1 Generating datasets from undirected networks

Search-and-score methods typically assume the presence of causality (*e.g.* in the form of a DAG model) in the underlying network that they aim at discovering. By contrast, constraint-based algorithms and the 3off2 approach can in principle ascertain the presence of causal relationships in the underlying graphical model. To evaluate the 3off2 algorithm on observational data related to undirected networks, Giulia Malaguti (former PhD student of Isambert's group, post-doc student at FOM Institute AMOLF) has generated large datasets by applying the Metropolis algorithm to an Ising model where the spins correspond to the nodes of an undirected network, and the pairwise interactions between the spins correspond to the edges of this network. Each edge $i-j$ is associated to a weight J_{ij}, a non-zero element of an adjacency matrix J that describes the direct pairwise coupling between the spins. Successive spin configurations are obtained by flipping each spin of the current configuration with a given probability p and by accepting or rejecting the resulting proposed configuration according to its Boltzmann probability. This provides a mean to generate large datasets corresponding to undirected networks with skeleton corresponding to the non zero elements of the matrix J.

One caveat of this data generating process is that the samples, corresponding to the successive configurations of the system, are correlated by construction (to obtain a significant acceptance rate in the Metropolis algorithm). As the 3off2 reconstruction method relies on the number of *independent* samples N to learn the skeleton and the orientations, it is necessary to implement a method to adjust the number N of correlated configurations to a smaller number N_{eff}, *i.e. effective* N, corresponding to the estimated number of independent samples.

Hence, defining the fluctuations of each variable k around its mean (over the whole dataset) as,

$$\sigma_k^i = s_k^i - \langle s_k^i \rangle_i$$
$$\sigma_k^{i+n} = s_k^{i+n} - \langle s_k^{i+n} \rangle_i$$

We then introduce the autocorrelation function $C_{exp}(n)$ defined as,

$$C_{exp}(n) = \frac{\frac{1}{N-n}\sum_{i=1}^{N-n}\frac{1}{K}\sum_{k=1}^{K}\sigma_k^i\sigma_k^{i+n}}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{k=1}^{K}{\sigma_k^i}^2}$$

and compare its decay with an exponential function $\exp(-n/R)$. This enables us to estimate an effective number N_{eff} of independent configurations as $N_{\mathit{eff}} = N/R$.

7.3.2 Reconstructing simulated undirected networks

The skeletons of the ALARM and INSURANCE networks have been chosen as models to set the non zero weights of the adjacency matrix J. In order to evaluate the 3off2 reconstruction method on the discovery of the absence of causality, without being impaired by errors in the learning of the skeleton, the weights of the adjacency matrix J have been set to the (conditional) mutual informations computed with the 3off2 method when reconstructing the oriented ALARM or INSURANCE networks from a large dataset. These conditional mutual informations indeed amount to the remaining direct interactions between pairs of variables once the indirect effects of all positively contributing variables have been removed from the mutual informations.

To reconstruct the *undirected-ALARM* and *undirected-INSURANCE* networks, datasets of size $N = 10^6$ have been generated. Tables 7.1 & 7.2 give the true positive (TP), false positive (FP) and false negative (FN) counts for the reconstructions obtained with the 3off2 method. When assuming that all the samples are independent ($N = 10^6$), the 3off2 method orients all the discovered edges. However, when adjusting for the number of samples, the 3off2 method rapidly indicates the absence of causality, and thus of orientations. Following the estimation method given in section 7.3.1, the estimated N_{eff} for *undirected-ALARM* and *undirected-INSURANCE* networks is approximately $130,000$ (Figs.7.11 & 7.12). From $N_{eff} = 100,000$ to $N_{eff} = 150,000$, the 3off2 method gives one of the best F-score for both undirected networks, and without any orientated edge. In fact, the 3off2 approach already performs well at $N_{\mathit{eff}} \leqslant 800,000$ for *undirected-ALARM* and $N_{\mathit{eff}} \leqslant 700,000$ for *undirected-INSURANCE* (Tables 7.1 & 7.2).

N_{eff}	TP	FP	FN	Prec.	Rec.	Fs.	%ort.
1,000,000	43	79	3	0.35	0.93	0.51	100%
900,000	43	74	3	0.37	0.93	0.53	100%
800,000	43	70	3	0.38	0.93	0.54	100%
700,000	43	57	3	0.43	0.93	0.59	100%
500,000	42	36	4	0.54	0.92	0.68	100%
250,000	38	8	8	0.83	0.83	0.83	**0%**
200,000	38	3	8	0.93	0.83	0.87	**0%**
150,000	38	0	8	1.00	0.83	0.90	**0%**
125,000	37	0	9	1.00	0.80	0.89	**0%**
100,000	37	0	9	1.00	0.80	0.89	**0%**
75,000	36	0	10	1.00	0.78	0.88	**0%**
50,000	36	0	10	1.00	0.78	0.88	**0%**
25,000	34	0	12	1.00	0.74	0.85	**0%**
10,000	34	0	12	1.00	0.74	0.85	**0%**

TABLE 7.1 **Evaluation of the** 3off2 **reconstructions of** *undirected-ALARM* **networks by** N_{eff}**.** The table gives the number of true positive (TP), false positive (FP) and false negative (FN) edges obtained when reconstructed the *undirected-ALARM* network with the 3off2 method (NML complexity). The precision (Prec.), recall (Rec.) and f-score (Fs.) evaluates the performance of the skeleton discovery. The %ort. gives the proportion of oriented edges within the learnt interactions.

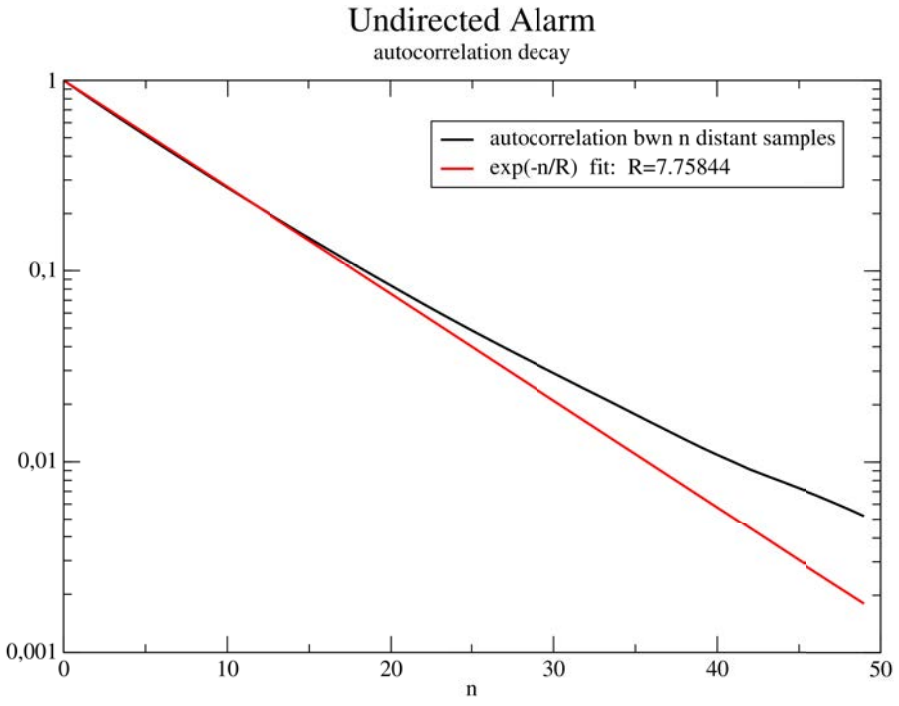


FIGURE 7.11 ***undirected-ALARM* autocorrelation decay.**

N_{eff}	TP	FP	FN	Prec.	Rec.	Fs.	%ort.
1,000,000	42	25	10	0.63	0.81	0.70	100%
900,000	42	23	10	0.65	0.81	0.72	100%
800,000	42	18	10	0.70	0.81	0.75	100%
700,000	41	15	11	0.73	0.79	0.76	100%
500,000	40	9	12	0.82	0.71	0.79	100%
250,000	37	2	15	0.95	0.71	0.81	**0%**
200,000	37	1	15	0.97	0.71	0.82	**0%**
150,000	37	0	15	1.00	0.71	0.83	**0%**
125,000	36	0	16	1.00	0.69	0.82	**0%**
100,000	35	0	17	1.00	0.67	0.80	**0%**
75,000	35	0	17	1.00	0.67	0.80	**0%**
50,000	35	0	17	1.00	0.67	0.80	**0%**
25,000	34	0	18	1.00	0.65	0.79	**0%**
10,000	30	0	22	1.00	0.58	0.73	**0%**

TABLE 7.2 **Evaluation of the** 3off2 **reconstructions of** *undirected-INSURANCE* **networks by** N_{eff}**.** The table gives the number of true positive (TP), false positive (FP) and false negative (FN) edges obtained when reconstructed the *undirected-ALARM* network with the 3off2 method (NML complexity). The precision (Prec.), recall (Rec.) and f-score (Fs.) evaluates the performance of the skeleton discovery. The %ort. gives the proportion of oriented edges within the learnt interactions.

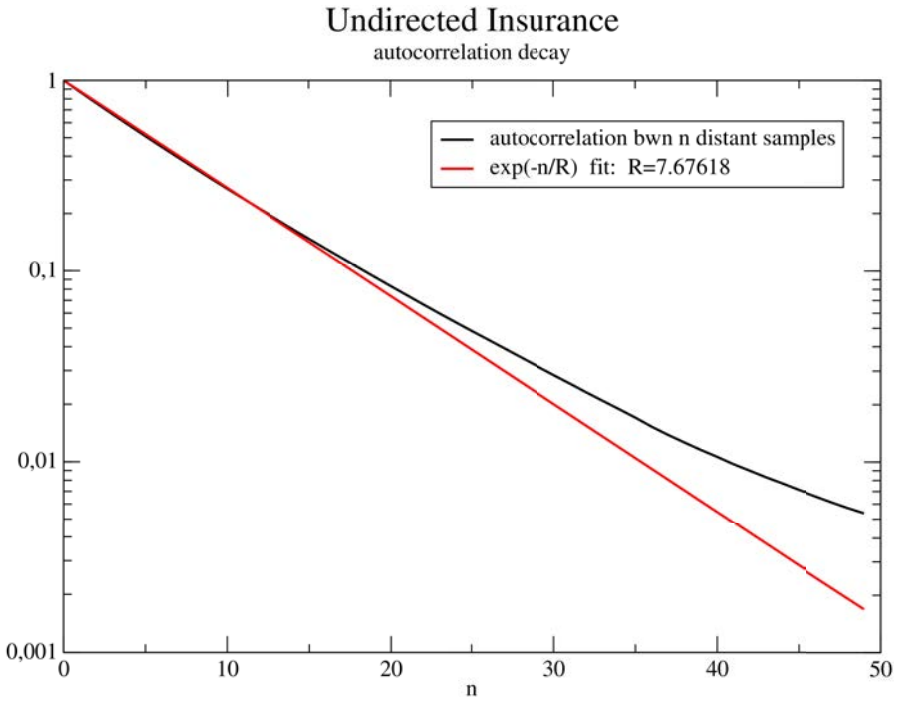


FIGURE 7.12 ***undirected-INSURANCE*** **autocorrelation decay.**

By contrast, as shown in Tables 7.3 & 7.4, the PC algorithm (PC-stable version using either the majority or the conservative orientation rule [Colombo and Maathuis, 2014, Ramsey et al., 2006]), requires an unusually small significance level

to provide fully undirected networks (Tables 7.3 & 7.4, $\alpha \leqslant 10^{-20}$ for *undirected-ALARM* and $\alpha \leqslant 10^{-7}$ for *undirected-INSURANCE*).

α	TP	FP	FN	Prec.	Rec.	Fs.	%ort.maj.	%ort.csv.
$1e^{-2}$	44	142	2	0.24	0.96	0.38	95%	45%
$1e^{-3}$	43	96	3	0.31	0.93	0.46	81%	45%
$1e^{-5}$	43	42	3	0.51	0.93	0.66	69%	43%
$\mathbf{1e^{-6}}$	42	27	4	0.61	0.91	0.73	35%	**0%**
$1e^{-7}$	42	17	4	0.71	0.91	0.80	7%	3%
$1e^{-10}$	40	5	6	0.89	0.87	0.88	4%	4%
$\mathbf{1e^{-20}}$	37	0	9	1.00	0.80	0.89	**0%**	**0%**
$1e^{-30}$	36	0	10	1.00	0.78	0.88	**0%**	**0%**
$1e^{-40}$	36	0	10	1.00	0.78	0.88	**0%**	**0%**

TABLE 7.3 **Evaluation of the PC-stable reconstructions of** *undirected-ALARM* **networks by** N_{eff}**.** The table gives the number of true positive (TP), false positive (FP) and false negative (FN) edges obtained when reconstructing the *undirected-ALARM* network with the PC-stable algorithm. The precision (Prec.), recall (Rec.) and f-score (Fs.) evaluates the performance of the skeleton discovery. The %ort. gives the proportion of oriented edges within the learnt interactions. The proportion of oriented edges when using either the majority or the conservative rule for the orientation step is given by %ort.maj. and %ort.csv. respectively.

α	TP	FP	FN	Prec.	Rec.	Fs.	%ort.maj.	%ort.csv.
$1e^{-2}$	47	64	5	0.42	0.90	0.58	97%	54%
$1e^{-3}$	43	33	9	0.57	0.83	0.68	89%	62%
$1e^{-5}$	41	16	11	0.72	0.79	0.75	75%	58%
$1e^{-6}$	41	9	11	0.82	0.79	0.80	82%	68%
$\mathbf{1e^{-7}}$	40	8	12	0.83	0.77	0.80	75%	**0%**
$\mathbf{1e^{-10}}$	39	1	13	0.98	0.75	0.85	**0%**	**0%**
$\mathbf{1e^{-20}}$	37	0	15	1.00	0.71	0.83	**0%**	**0%**
$\mathbf{1e^{-30}}$	36	0	16	1.00	0.69	0.82	**0%**	**0%**
$\mathbf{1e^{-40}}$	35	0	17	1.00	0.67	0.80	**0%**	**0%**

TABLE 7.4 **Evaluation of the PC-stable reconstructions of** *undirected-INSURANCE* **networks by** N_{eff}**.** The table gives the number of true positive (TP), false positive (FP) and false negative (FN) edges obtained when reconstructing the *undirected-INSURANCE* network with the PC-stable algorithm. The precision (Prec.), recall (Rec.) and f-score (Fs.) evaluates the performance of the skeleton discovery. The %ort. gives the proportion of oriented edges within the learnt interactions. The proportion of oriented edges when using either the majority or the conservative rule for the orientation step is given by %ort.maj. and %ort.csv. respectively.

These reconstructions of undirected networks performed by the 3off2 method support its ability to uncover the absence of causality from observational datasets. The PC algorithm can also discover, for these benchmark networks, the absence of orientations. Yet, the PC algorithm requires a smaller significance level as compared to the values typically used in the literature, making it difficult to decide whether there is actually an absence of orientations, or a non detection of orientations.

Applications & Perspectives

Chapter 8

Interplay between genomic properties on the fate of gene duplicates with the 3off2 method

The mechanism underlying the emergence and the expansion of 'dangerous' gene families is an evolutionary oddity from a natural selection perspective. While the maintenance of 'essential' genes[1] is ensured by their lethal null mutations, the expansion of 'dangerous' gene families implicated in cancer and other severe genetic diseases remains puzzling. These genes, albeit critical to normal cellular functions, can be referred to as 'dangerous' owing to their susceptibility to dominant deleterious mutations leading to diseases. Surprisingly, 'dangerous' genes have been duplicated more than others in the course of vertebrate evolution. What could have been the evolutionary causes responsible for such striking expansion? Have the gene susceptibility to dominant deleterious mutations played a driving role?

In this chapter, I present converging evidence supporting the hypothesis that 'dangerous' genes in the human genome have been preferentially retained after two rounds of whole-genome duplication (WGD) dating back from the onset of jawed vertebrates, some 500MY ago [Ohno, 1970, Putnam et al., 2008]. I further demonstrate that the retention of many WGD-duplicated genes, so-called 'ohnologs'[2], suspected to be dosage balanced[3], is in fact *indirectly mediated* by their susceptibility to deleterious mutations. In addition I introduce a somewhat counterintuitive

[1] Refer to Chapter 1, footnote[4]
[2] Refer to Chapter 1, footnote[3]
[3] The 'dosage-balance' hypothesis posits that the ohnologs are retained because their interactions with protein partners require to maintain balanced expression levels throughout evolution.

yet simple evolutionary model accounting for this enhanced retention of 'dangerous' ohnologs. Finally, I discuss the networks of genomic properties reconstructed using the 3off2 method that support the proposed evolutionary model.

8.1 Enhanced retention of ohnologs

8.1.1 Enhanced retention of 'dangerous' ohnologs

We performed a data mining analysis that revealed a strong correlation between the retention of human ohnologs and their reported susceptibility to deleterious mutations [Singh et al., 2012]. The statistics computed from this data mining analysis were based on the 20,506 protein coding genes of the human genome (Ensembl release 61). 35% of these genes can be traced back to the two WGD events at the onset of jawed vertebrates [Makino and McLysaght, 2010]. We obtained, in particular, cancer gene datasets from multiple databases, including COSMIC [Forbes et al., 2008] and CancerGenes [Higgins et al., 2006], other genetic disease genes from OMIM and dominant negative[4] genes and genes with autoinhibitory protein folds[5] from literature search.

It appears that the human genes associated with the occurence of cancer and other genetic diseases (8,095) have retained significantly more ohnologs than expected by chance, 48% *versus* 35% (48%; 3,844/8,095; $P = 1.3 \times 10^{-129}$, χ^2). Furthermore, correlations are clearly enhanced when the analysis is restricted to genes with direct experimental evidence of dominant deleterious mutations, such as for genes with dominant negative mutants (61%; 292/477; $P = 3.9 \times 10^{-34}$, χ^2).

As shown later in section 8.4, networks of genomic properties reconstructed with the 3off2 method infer a direct causal effect between the properties 'Cancer' (*i.e.* gene prone to mutations involved in cancer) or *'Dom_Neg'* (*i.e.* gene prone to dominant negative mutations) and *'Ohnolog'* (*i.e.* gene having retained a copy from a whole genome duplication), indicating that a gene prone to dominant negative mutations or participating in a cancer is more likely to retain copies from WGD events (Figs. 8.3 & 8.4).

[4] Dominant negative phenomena occur in diploid organism when a mutated allele adversely interferes with the functional allele.

[5] Autoinhibitory proteins are multidomain proteins where the domain having the functional property is inhibited by interaction with another part of the protein through folding. Conformational changes can for instance be triggered by interactions with other proteins.

8.1.2 Retained ohnologs are more 'dangerous' than dosage balanced

We have also found [Singh et al., 2012, 2014] that the reported autosomal dominant disease genes (440) have experienced stronger retention biases in ohnologs (59%; 261/440; $P = 1.7 \times 10^{-27}, \chi^2$) than the reported autosomal recessive diseases genes (598) (37%; 221/598; $P = 0.24$, χ^2). Besides, the human orthologs of mouse genes reported as being 'essential' genes (2729) from large scale null mutant studies in mouse (5956 genes tested) exhibit a small enrichment of ohnologs, 56 *vs* 54% (56%; 1537/2729; $P = 3.8 \times 10^{-3}$, χ^2), where 54% (3190/5656) is the global proportion of ohnologs among the genes tested in mouse for null mutation.

These results, in association with complementary statistics not shown here [Singh et al., 2012], suggest that the retention of ohnologs is more strongly related to their 'dangerousness' than their 'essentiality' or 'dosage-balance' [6]. As shown in section 8.4, the networks of genomic properties obtained from these observational data with the 3off2 method infer no direct relationship between the properties *'Essential'* or *'Complex'* and *'Ohnolog'*.

8.2 Going beyond simple correlations in genomic data

A number of studies have shown that many genomic properties are, to some extent, correlated. Among them, gene essentiality, functional ontology, expression level, divergence rates, etc, have been widely considered. Yet, many of these correlations, suggesting direct statistical association, could in fact be mediated through the indirect effect of third properties and only provide partial insights on complex biological systems. To go beyond such simple statistical associations and quantify the direct and indirect effects of genomic properties, I have performed Mediation analyses, following the approach of Pearl [Pearl, 2001, 2012]. In particular, I aimed at disentangling the relative effects of deleterious mutations and dosage balance constraints on the retention of ohnologs.

8.2.1 The Mediation framework in the context of causal inference

The Mediation analysis, typically used in social sciences and epidemiology, aims at uncovering causal pathways along which changes in multivariates properties are transmitted from a cause, X, to an effect, Y. In other words, the Mediation

[6] Refer to footnote 3

analysis assesses the importance of a mediator, M, in transmitting the indirect effect of X on the response Y. In the Mediation framework, the average direct effect, $DE_{xx'}$, is defined as the change that Y would experience if x could be changed to x' while keeping M fixed. Similarly, the average indirect effect, $IE_{xx'}$, is defined as the change that Y would experience if the mediator could be changed from $m(x)$ to $m(x')$ while keeping X to its original value x (Fig. 8.1 & Table 8.1, ***A***.) [Pearl, 2012]. In fact, the counterfactual expressions used to quantify these effects formally decouple the direct and indirect conditions on X, seen by the outcome Y.

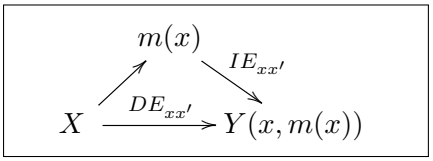


FIGURE 8.1 **Mediation diagram.** This diagram shows the direct effect, $DE_{xx'}$, of changing X from x to x' onto the outcome Y, and the indirect effect, $IE_{xx'}$, mediated by $m(x)$.

In the framework of Bayesian statistics, the Mediation formulae can be expressed as in Table 8.1, ***B***., where $E\left(Y|x,m\right)$ denotes the expectation value of Y given X and M, and $P\left(M|x\right)$ denotes a value of M drawn from a distribution P conditionally on X. Owing to non-linear couplings, $DE_{xx'}$ and $IE_{xx'}$ usually do not sum to the total effect $TE_{xx'}$. However, the proportions $DE_{xx'}/TE_{xx'}$ and $IE_{xx'}/TE_{xx'}$ can be interpreted in terms of sufficient *versus* necessary contributions to $TE_{xx'}$ [Pearl, 2012].

Effects	***A***. Shorthand notation	***B***. Bayesian Framework
$DE_{xx'}$	$Y(x',m(x)) - Y(x,m(x))$	$\sum_m [E(Y\|x',m) - E(Y\|x,m)]P(M\|x)$
$IE_{xx'}$	$Y(x,m(x')) - Y(x,m(x))$	$\sum_m E(Y\|x,m)[P(M\|x') - P(M\|x)]$
$TE_{xx'}$	$Y(x',m(x')) - Y(x,m(x))$	$E(Y\|x') - E(Y\|x)$

TABLE 8.1 **Mediation analysis formulae.** This table gives the formulae corresponding to the direct effect, $DE_{xx'}$, and the indirect effect, $IE_{xx'}$, mediated my $m(x)$.

If X, M, and Y are binary variables, expectation values of Y, and values of M drawn from a conditional distribution P, can be estimated from the counts of the different possible triples of values for (X,M,Y). Let's define n_{xmy} as the number of (X,M,Y) triples where $(X,M,Y) = (x,m,y)$. Then, $E\left(Y|X=0,M=0\right)$ amounts to $\left(\frac{n_{001}}{n_{000}+n_{001}}\right)$. Similarly, $P\left(M|X=0\right)$ can be estimated by $\left(\frac{n_{010}+n_{011}}{n_{000}+n_{001}+n_{010}+n_{011}}\right)$.

8.2.2 The Mediation analysis applied to genomic data

I have applied the causal inference methods to the 20,506 protein-coding genes in human. In particular, I focused on the disentanglement of direct and indirect effects favoring the retention of duplicated genes after the two rounds of whole-genome duplication (*WGD*) events that occured at the onset of jawed vertebrates, some 500MY ago [Ohno, 1970, Putnam et al., 2008]. More precisely, I quantified the effect of the dosage balance constraints ('dosage.bal.', as X or M) and the effect of the gene susceptibility to deleterious mutations ('delet.mut.', as X or M) on the retention of ohnologs ('ohno.', as Y).

With respect to our data mining analysis [Singh et al., 2012], I have divided the set of human protein coding genes into two categories: (***1***) 'dosage.bal.' (4,003 genes), including 3,814 'complex' genes[7], and (***2***) 'delet.mut.' (7,227 genes), including 6,917 'cancer' genes, 440 'dominant disease' genes[8], 477 'dominant negative' genes and 461 'autoinhibited' genes. Following Pearl's formulae (Table 8.1, ***B***), I quantified the direct and indirect effects of 'dosage.bal.' and 'delet.mut.' properties on the ohnolog retention, 'ohno.'. The estimated sufficient (direct) contributions or necessary (indirect) contributions (Table 8.2) demonstrate that the retention of human ohnologs is more strongly caused by their susceptibility to deleterious mutations, either as *direct* or *indirect* factor, than their interactions within multi-protein complexes.

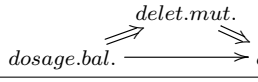
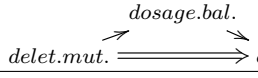
Mediation Diagram	TE	DE/TE	IE/TE	Non-linear Coupling
	9.0×10^{-2}	29.4%	**74.1%**	3.5%
	2.3×10^{-1}	**98.3%**	2.6%	1.0%

TABLE 8.2 **Mediation analysis results.** This table gives direct and indirect estimations between three genomic properties, namely the dosage-balance constraint, *'dosage.bal.'*, the susceptibility to deleterious mutations, *'delet.mut.'* and the retention of ohnologs, *'ohno.'*.

In fact, as will be shown in section 8.4, the networks of genomic properties obtained with the 3off2 method infer no direct relationship between the properties 'Complex' and 'Ohnolog', suggesting that the propensity of a gene to retain its copies from WGD events is not a direct consequence if its involvement in multi-protein complexes (Figs. 8.3 & 8.4).

[7] Genes that participate in multi-protein complexes.
[8] Dominant disease genes are genes prone to dominant deleterious mutations.

Recent genome wide analyses have shown that ohnologs, and SSD[9] (Small Scale Duplication) or CNV[10] (Copy Number Variation) exhibit antagonist retention patterns. Besides, ohnologs in the human genome have experienced fewer SSD than 'non-ohnolog' [Makino and McLysaght, 2010]. As the SSD constraint could biased our analysis, I have also considered only the $8,215$ human genes without SSD nor CNV duplicates. The direct effect of 'dosage.bal.' tends then to oppose the retention of ohnologs (-33.4%), while the indirect effect, mediated by 'delet.mut.', is even more pronounced (130.5%).

Interestingly, we will show in section 8.4 that the 3off2 method recovers this antagonist patterns, as can be seen in Figs. 8.3 & 8.4, that exhibit a negative causal effect from 'Ohnolog' towards 'recentSSD'.

8.3 Retention of 'dangerous' ohnologs through a counter-intuitive mechanism

These findings have led the Isambert's group to a new evolutionary model describing a non-adaptive mechanism through which 'dangerous' ohnologs could have been retained during the course of evolution. Fig. 8.2 examines all possible evolutionary scenarios following a WGD event in the genome of one or a few individuals in an initial population [Affeldt et al., 2013, Singh et al., 2012].

A critical feature of WGD events compared to SSD events is their coupling to subsequent speciation events. Following the 'duplication' and the WGD-induced 'speciation', there could be three possible scenarios under 'mutation/selection' process, leading to fixation (Fig. 8.2) or loss of individual ohnologs:

(A) Retention of ohnologs by spread and fixation of beneficial mutations in the post-WGD population by positive selection

(B) Loss of one copy by fixation of neutral non-functionalizing mutations by drift, also leading to further speciation in the post-WGD population

(C) Elimination of individual(s) harboring dominant deleterious mutations, leading to a post-WGD population composed of surviving individuals still holding the (non-deleterious) functional copies of the 'dangerous' ohnologs

[9] Small scale duplicates correspond to duplicated regions ranging from a few base pairs to a large genomic segment.

[10] Copy number variations are a form of small scale duplications, and correspond to variants within a population where individual human can have different (number of) copies of a part of their genome.

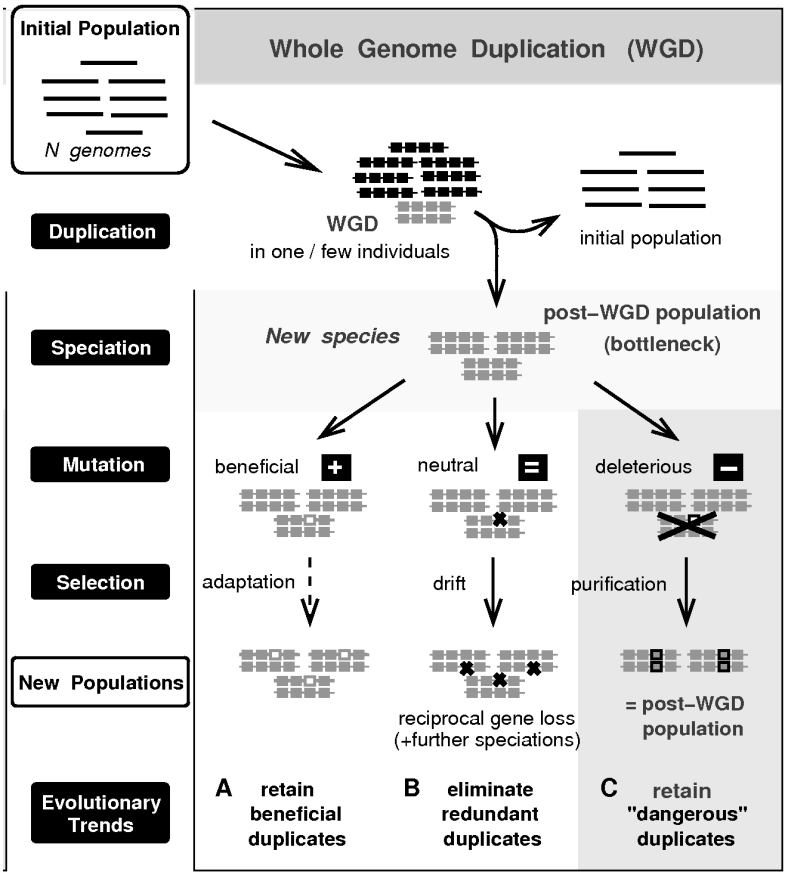


FIGURE 8.2 **Evolutionary trends of duplicated genes following a WGD.** Horizontal lines represent the genome of different individuals. Square blocks symbolize the genes, duplicated (grey) or not (black). Black crosses highlight the loss of one gene (small crosses) or the elimination of an individual (larger crosses), while bordered square blocks emphasize retained mutated copies. See main text for a detailed discussion of the three mutation/selection scenarios (A-C) and resulting evolutionary trends. From [Affeldt et al., 2013]

Note, in particular, that this counterintuitive model for the retention of 'dangerous' ohnologs, hinges on *(i)* the *speciation* event concomitant to WGD, owing to the difference in ploidy between pre- and post-WGD individuals and *(ii) autosomal dominance* of deleterious mutations leading to purifying selection (C). This model is also supported by deterministic and stochastic simulations done in the Isambert's group [Malaguti et al., 2014].

Although the retention of ohnologs remains stochastic, the susceptibility of 'dangerous' genes to dominant deleterious mutations significantly increases the odds of ohnolog retention through path (C), Fig. 8.2. By contrast, the bottleneck due to small WGD-population size limits the efficacy of adaptation that could lead eventually to fixation of beneficial mutations through path (A), Fig. 8.2. Therefore, as suggested by our data mining analysis [Singh et al., 2012], most 'dangerous'

ohnologs are retained by purifying selection (*i.e.* path (C)) due to their susceptibility to dominant deleterious mutations. In other words, purifying selection has minimized the rapid loss of 'dangerous' genes in post WGD population (Fig. 8.2) and the ensuing sub-functionalization has ensured their long term survival in post-WGD species [Malaguti et al., 2014].

8.4 Reconstructing the interplay of genomic properties on the retention of ohnologs

8.4.1 Genomic properties of interest

The 3off2 reconstruction method has been used to learn causal graphical models among several genomic properties associated with human protein coding genes and for which information has been computed in our group or collected from the literature. As I was primarily interested in disentangling the direct and indirect effect of the susceptibility to dominant deleterious mutations, and dosage balance constraints on the retention of ohnologs, I have first considered the genomic properties from (1) to (9) listed in Table 8.3. These properties also include SSD in order to uncover the antagonist pattern retention between SSD and WGD duplicates. Then, other properties, such as the copy number variation, the expression level or the degree of conservation of the protein sequence, have been taken into account, leading to twelve properties, Table 8.3.

As direct information of human essential gene is not available, the 'essential' data correspond either to essentiality in mouse orthologous[14] genes (Mouse Genome Informatics (MGI) database [Eppig et al., 2012]). The essentiality of human protein coding genes could be assessed for only $6,436$ genes out of $20,415$. Yet, as previously detailed in section 6.4.3, the 3off2 approach can deal with missing data by computing the 2-point and 3-point information terms solely on the observed variables, without impairing the other calculations.

The Ka.Ks property indicates the conservation of the gene sequence. Indeed, the Ka/Ks ratio measures the proportion of nonsynonymous substitutions (Ka) to the proportion of synonymous substitutions (Ks). I have shown in [Singh et al., 2012] that ohnologs exhibit statistically lower Ka/Ks ratios, which provided direct evidence of strong conservation, consistent with a higher susceptibility of ohnologs to

[14]Genes in the genome of two different species sharing the same ancestor gene.

Property	Description	datapoints
(1) Ohnolog	Copies retained from WGD events	20,415
(2) recentSSD	Copies retained from SSD events after the two rounds of WGD at the origin of vertebrates	20,415
(3) Cancer	Prone to mutations implicated in cancer	20,415
(4) Disease	Prone to mutations implicated in diseases	20,415
(5) Haploinsuf	Haploinsufficient[11]	20,415
(6) Dom_Neg	Prone to mutations implicated in dominant negative[12] diseases	20,415
(7) Essential	Essential	**6,436**
(8) Autoinhib	Protein with autoinhibitory folds[13]	20,415
(9) Complex	Involved in multi-protein complexes	20,415
(10) CNV	Presence of copy number variations	20,415
(11) ExpressionLevel	Expression level	**13,425**
(12) Ka.Ks	Conservation measure of the protein sequence	**15,508**

TABLE 8.3 **Genomic properties for the reconstruction of causal graphical models in the fate of gene duplicates.** Twelve genomic properties participating in the fate of genome duplicates in the course of vertebrates evolution. The status of each property has been identified for all human protein coding genes, except for 'Essential', 'Ka.Ks' and 'ExpressionLevel'. See main text for complementary information.

deleterious mutations. Genes for which Ka/Ks ratio could not be calculated were discarded from the analysis, leading to 15,508 datapoints for this property. The remaining continuous Ka/Ks values have been binarized 0/1 with the threshold at 75% quantile across all human protein coding genes. Thus, the Ka.Ks property used in the 3off2 reconstructions (Fig. 8.3 & 8.4) corresponds to 'high divergence of the sequence' (0[no]/1[yes]).

Finally, gene expression values were downloaded from BioGPS [Wu et al., 2009] and correspond to the median values of expression across 78 tissues/cell types (13,372 genes in total).

8.4.2 Non-adaptive retention mechanism supported by the 3off2 causal models

The interactions between the first 9 genomic properties, as recovered by the 3off2 method, are given in Fig. 8.3. The colors indicate the strength of the relationships (from 'weak' [light blue/yellow/light grey] to strong [dark blue/red/dark grey]). The arrowheads indicate the direction of the effect, and the colors provide also the sign, with warm colors indicating a positive effect and cold colors indicating a negative effect. Finally, the non oriented edges suggest an absence of cause-effect

mechanism, while the bi-directed edges indicate a common unobserved cause from a *latent variable* (section 6.4.2).

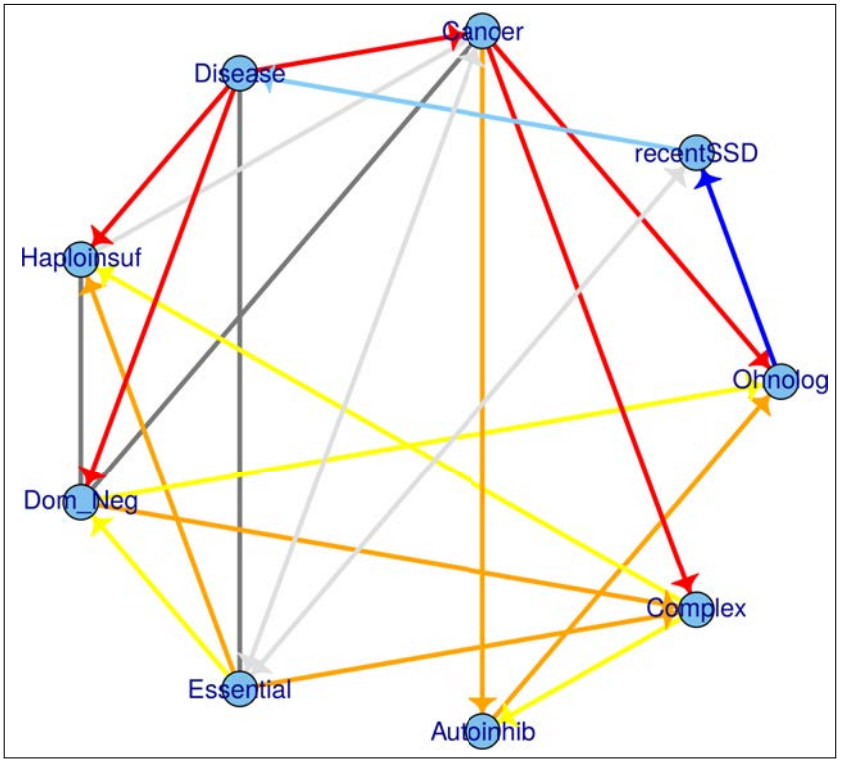


FIGURE 8.3 **Causal network of 9 genomic properties reconstructed with the** 3off2 **method.** The colors indicate the strength of the relationships (from 'weak' [light blue/yellow/light grey] to strong [dark blue/red/dark grey]) and provide also the sign, with warm colors indicating a positive effect and cold colors indicating a negative effect. The arrowhead gives the direction of the effect. The non oriented edges suggest an absence of cause-effect mechanism, while the bi-directed edges indicate a common unobserved cause. See main text for detailed on the properties.

One of the first conclusions that can be drawn from Fig. 8.3 is that the 3off2 method does not recover any direct relationship between 'Complex' and 'Ohnolog'. The former indicating the involvement of the gene in multi-protein complexes, it embodies dosage-balance constraints. Thus, this graphical model confirms that the retention of ohnolog does not primarily rely on this alternative hypothesis by [Makino and McLysaght, 2010]. A second interesting result is that 'Cancer', 'Dom_Neg' and 'Autoinhib' properties directly point toward the 'Ohnolog' node, supporting our hypothesis that the susceptibility to (dominant) deleterious mutations participates to the retention of copies from the 2 rounds of WGD events that occurred at the origin of vertebrates. One can also recover from Fig. 8.3 the antagonist retention pattern between genes prone to SSD duplicates and genes prone to WGD duplicates. Indeed, the edge linking 'Ohnnolog' and 'recentSSD' displays

an anti-correlation effect, with the edge direction consistent with the temporal retention of 'recentSSD' after the 2R-WGD in early vertebrates.

The graphical model given in Fig. 8.4 corresponds to the reconstructed causal graphical model when adding the properties (10) to (12) from Table 8.3. Note that the previous conclusions from Fig. 8.3 still hold, highlighting the robustness of our approach to the inclusion of additional properties (although the orientation between 'Ohnolog' and 'Autoinhib' is inverted).

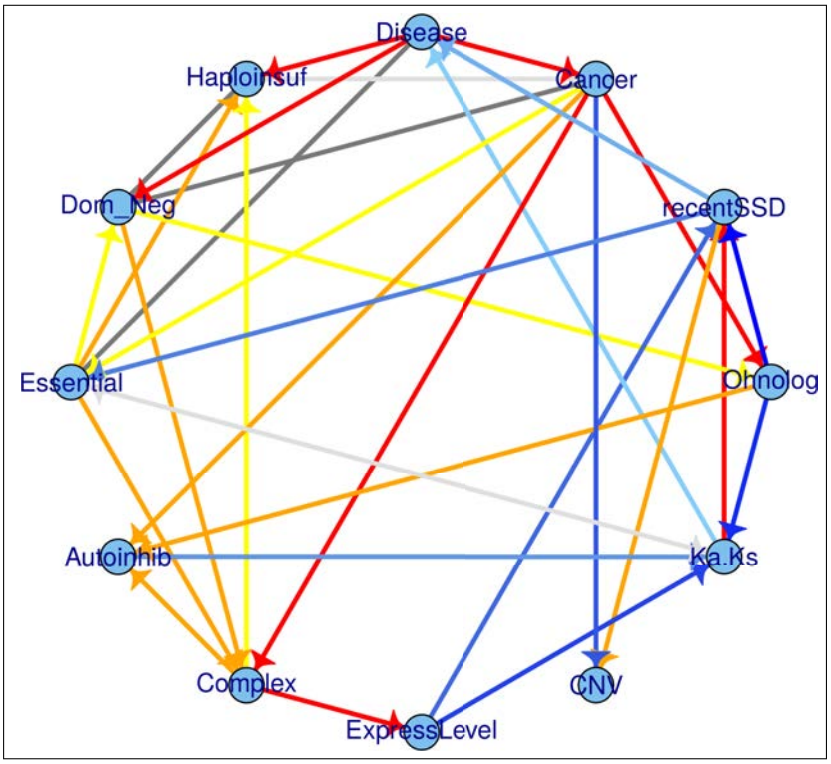


FIGURE 8.4 **Causal network of 12 genomic properties reconstructed with the** 3off2 **method.** The colors indicate the strength of the relationships (from 'weak' [light blue/yellow/light grey] to strong [dark blue/red/dark grey]) and provide also the sign, with warm colors indicating a positive effect and cold colors indicating a negative effect. The arrowhead give the direction of the effect. The non oriented edges suggest an absence of cause-effect mechanism, while the bi-directed edges indicate a common unobserved cause. See main text for detailed on the properties.

Interestingly, the Ka.Ks property is *negatively* pointed by 'Ohnolog', confirming our previous comparative genomic results on the conservation of ohnolog sequences. Besides, the cause-effect relationship between 'Ka.Ks' and 'recentSSD' is positive, supporting widely acknowledged results on the necessary positive selection (*i.e.* with higher Ka/Ks ratios) of SSD copies. We also note that genes with recent SSD tend to be less prone to be essential and implicated in disease as their SSD copy can buffer to some extent the deleterious effects of loss-of-function mutations.

All in all, these results obtained with the 3off2 reconstruction approach support the non-adaptive retention mechanism of ohnologs that have been proposed during the first part of my thesis [Affeldt et al., 2013, Malaguti et al., 2014, Singh et al., 2012]. This global overview provides valuable insights on the interplay between the genomic properties that influence directly or indirectly the fate of duplicates during the evolution enables. It also provides the causal graphical model that is needed to perform Mediation Analyses and quantify the causal relationships between these genomic properties.

Chapter 9

Reconstruction of zebrafish larvae neural networks from brain-wide *in-vivo* functional imaging

Common neural processes can span over several brain regions and involve large scale ensemble of neurones. Getting a comprehensive understanding of the brain functional networks thus requires the ability of acquiring and processing datasets corresponding to large fractions of the brain. Several imaging techniques have been proposed and widely used, such as two-photon or confocal microscopy, however their acquisition speed limitation has only allowed for the exploration of a small part of the brain so far. Recently, the Selective-plane Illumination Microscopy (SPIM) has provided images, at the level of the individual neuron and for large brain areas, allowing for such wide-brain analysis.

In these chapter, I report and discuss some network reconstructions obtained with the 3off2 method when applied to biological datasets generated from single-cell resolution images acquired with the advanced SPIM imaging technology. These datasets correspond to 2D layers of a zebrafish larvae brain at different depths (z-plane images), and have been provided by the research group of Dr. Georges Debrégeas and Dr. Raphaël Candelier (*Zebrafish behavior and calcium imagery* research group, Jean Perrin laboratory, UPMC, Paris VI). Sébastien Wolf performed the experiments using a SPIM imaging system and processed the acquired images of the spontaneous brain activity of zebrafish in their larval stage. The binary datasets obtained from this single-cell resolution imaging system were used as input for the 3off2 method to reconstruct functional networks corresponding to the

spontaneous brain activity of zebrafish larvae brain [Panier et al., 2013]. As shown in the section 9.2, the 3off2 reconstruction method could confirm already known communication patterns between large areas corresponding either to sensory or motor structures.

9.1 The zebrafish calcium functional imaging

9.1.1 The calcium functional imaging

Micro-arrays of electrodes can be used to measure the action potentials of a few hundred neurons at the same time with a high temporal resolution. Yet, this method only provides information on small areas, and cannot capture long range communications throughout the whole brain. By contrast, calcium imaging is a non-invasive technique that can optically monitor the fluctuations of Ca^{2+} concentration in a greater number of cells [Yuste and Katz, 1991]. These Ca^{2+} fluctuations reflect the action potentials that affect the cells during the spiking process as the sudden depolarization during a spike corresponds to the opening of charged ion channels within the cellular membrane. If the cells contain a calcium sensitive probe, such as GCaMP3 (a GFP protein associated with a calcium sensitive Calmoduline), a fluorescent signal can be optically recorded [Tian et al., 2009, 2012].

9.1.2 The zebrafish as vertebrate model

The zebrafish is a small vertebrate (typically $200 \times 500 \times 1000 \mu m$) that is well suited for functional calcium imaging. Its brain, which already contains tens of thousands of neurons in a larvae of a few days, is able to respond to several stimuli, like light or scent, at an early developmental stage when the whole body is transparent. In particular, the Debrégeas group has applied calcium functional imaging to observe neuronal cells of genetically modified zebrafish expressing GCaMP3 [Tian et al., 2009, 2012]. This calcium indicator provides a fluorescent signal related to neural spikes.

Fig. 9.1, adapted from [Portugues et al., 2014], highlights known functional areas in the larval zebrafish brain. As reported in previous studies, the regions of interest (ROIs) within the zebrafish brain display a clear symmetry, except for the habenula (Hab) region, which is known to have pronounced anatomical asymmetry and have

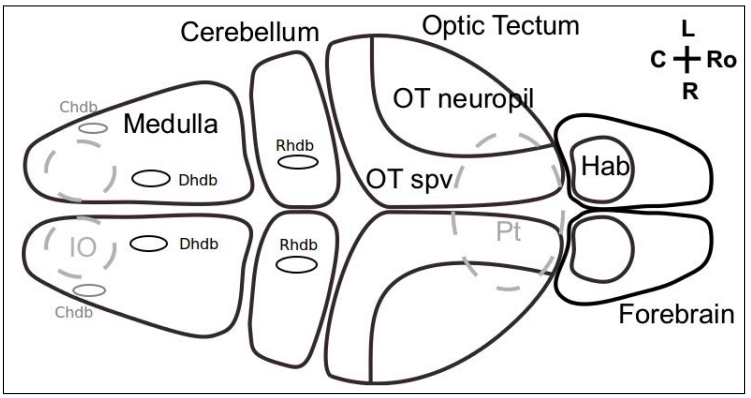


FIGURE 9.1 **Scheme of main brain areas in zebrafish larvae.** This scheme provides a dorsal view of zebrafish larvae brain. OT, optic tectum; OT spv, optic tectum stratum periventriculare; IO, inferior olive; Pt, pretectal area; Hab, habenula; Rhdb, rostral hindbrain; Chdb, caudal hindbrain; Dhdb, dorsal hinbrain. The gray dashed regions indicate more ventrally located areas.

predominant responses on the left side [Bianco and Wilson, 2009, Portugues et al., 2014].

Previous studies have shown that the activity in regions involved in sensory processing (*e.g.* tectal neuropil, pretectum (Pt) or inferior olive; Fig.9.1) tend to temporally precede the activity in the areas related to motor structures (*e.g.* rostral hindbrain (Rhdb) in the cerebellum region, ventral/dorsal hinbrain (Chdb/Dhdb) in the medulla region; Fig.9.1) [Portugues et al., 2014]. This temporal patterns are further discussed in section 9.2, in relation with the 3off2 network reconstructions.

9.1.3 The SPIM technique

9.1.4 The experimental setup

The Selective-Plane Illumination Microscopy (SPIM) [Holekamp et al., 2008, Keller et al., 2010, Panier et al., 2013] is a scanning laser sheet technology that performs an *optical sectioning* and collect images with a camera whose optical axis is orthogonal to the illumination plane. This technology allows the simultaneous observation of more neurons than with classical confocal or two photon microscopy, as different areas within the same z-plane are exposed for a much longer time on average. It has been shown[Panier et al., 2013] that the SPIM imaging technology reaches about 20-folds increase in the speed of acquisition, *i.e.* number of recorded neurons times acquisition rate, as compared with classical point-scanning imaging techniques (PSM). Specifically, PSM systems can acquire approximately one

thousand neurons at 5Hz, while SPIM imaging technology reaches about $5,000$ at 20Hz. The figure 9.2, taken from [Panier et al., 2013], gives an overview of the experimental settings available at the Jean Perrin laboratory.

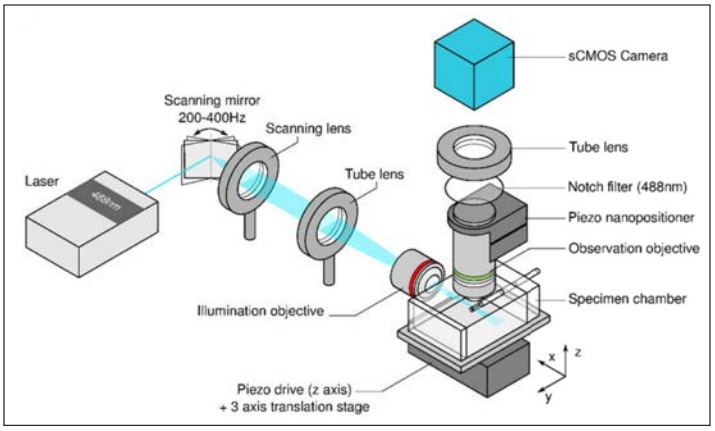


FIGURE 9.2 **Experimental setup of the SPIM-based functional imaging system (Jean Perrin laboratory, CNRS-UPMC).** Schema of the experimental SPIM setup at the Jean Perrin laboratory, G. Debregeas & R. Candelier team. [Panier et al., 2013].

As can be seen in Fig. 9.2, the zebrafish larvae is kept paralysed in a low-melting temperature agarose solution within a tank that can be vertically translated *via* a piezo-positioner, allowing observations at different depths of the brain. The laser light sheet enables the capture of fluorescent signals emitted by the neurons and provides an image which is formed on the sCMOS sensor. This image undergoes a series of manual and automated filtering to discard rare remaining movements of the paralysed larvae, remove high frequency noise and most importantly, correctly identify the activity of each individual neuron.

9.1.5 From fluorescent signal to neuron activity

The fluorescent signal formed on the sCMOS camera of the SPIM imaging system (Fig. 9.2), and the actual electrical signal differ by their amplitude, duration and shape. While the amplitude of the electrical signal is typically of about 100 mV, all fluorescent signals do not share the same amplitude. Besides, a spike lasts about 1ms while the relaxation of the fluorescent signal may reach up to 2s. Finally, instead of a spike, the fluorescent signal follows a double exponential law, $f(t) \propto e^{\lambda_r} + e^{-\lambda_d}$, where λ_r and λ_d are two characteristic times that depend on the calcium indicator (Fig. 9.4, blue curve; Results provided by Sébastien Wolf, Debrégeas research group). For each neuron, the fluorescence signal is not directly considered,

but instead compared to the cell baseline fluorescence, b_{slow}. Hence, the considered signal is $\Delta F/F = (F(t) - b_{slow})/b_{slow}$.

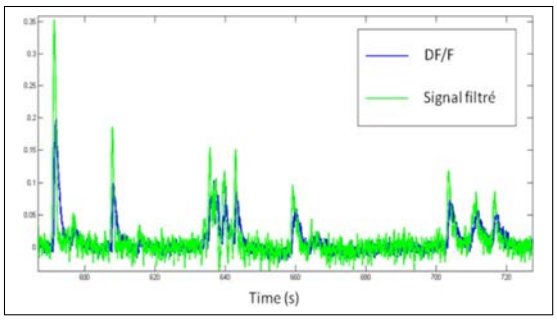


FIGURE 9.3 Activity of a zebrafish neuron (blue), and the corresponding filtered signal using a Wiener filter (green).

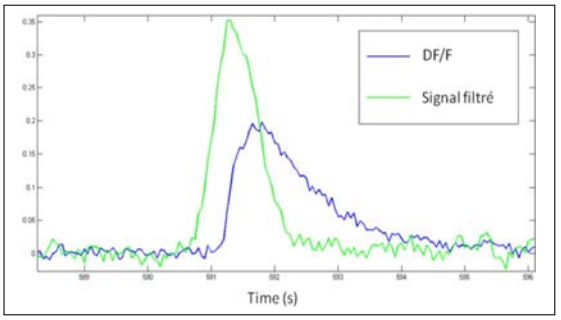


FIGURE 9.4 Magnification of figure 9.3

The remainig step is to extract the action potentials from the $\Delta F/F$ quantity. This extraction is a difficult task as it has to cope with many parameters such as the noise within the signal, the temporal resolution of the imaging system or the characteristics of the calcium indicators. A very simple extraction would consist in applying a threshold to the $\Delta F/F$ fluorescent signal, however more efficient methods that are based on deconvolution (*i.e.* Wiener filter) have also been proposed to extract the trains of action potentials [Holekamp et al., 2008]. Yet, the peaks of activity in the filtered signal still remain large, and hence the temporal resolution is low (Fig. 9.4, green curve). Thus, the group of G. Debrégeas developed another filtering method, inspired from [Ramdya et al., 2006], and based on the specific shape of the fluorescent signal, *i.e.* a rapid increase followed by a slow decrease (Fig. 9.4, blue curve). This processing, that relies on a Wiener filter and on a recursive procedure, produces good results on a range of tests performed in the Jean-Perrin laboratory.

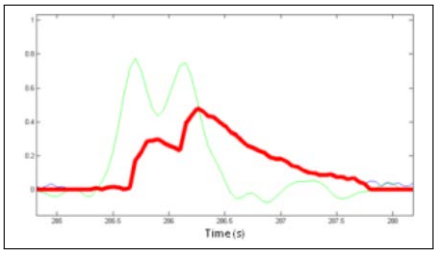


FIGURE 9.5 Action potential train of a zebrafish neuron (green), and the corresponding fluorescent signal (red).

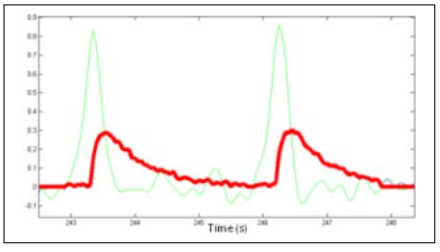


FIGURE 9.6 Succession of two temporally separated action potentials of a zebrafish neuron (green), and the corresponding fluorescent signal (red).

As can be seen in Fig. 9.5 (results provided by Sébastien Wolf, Debrégeas research group), the filter is able to recover individual peaks from a fluorescent signal, even when action potentials from a same train get superimposed. The method

proposed by G. Debrégeas correctly extracts temporally closed peaks, as shown in the Fig. 9.7 (results provided by Sébastien Wolf, Debrégeas research group).

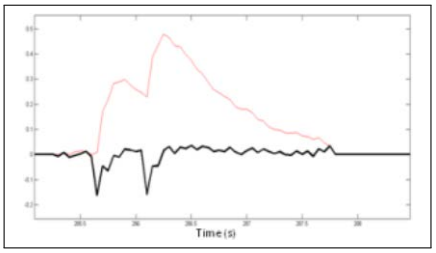


FIGURE 9.7 Fluorescent signal of a zebrafish action potential train (red) and the corresponding filtered signal (black) when the filter from the Jean-Perrin Laboratory is applied.

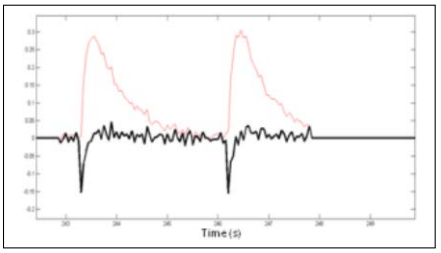


FIGURE 9.8 Fluorescent signal of two temporally separated action potentials of a zebrafish neuron (red) and the corresponding filtered signal (black) when the filter from the Jean-Perrin Laboratory is applied.

In section 9.2, the datasets used by 3off2 correspond to a binary version of the (downward) peak of Fig. 9.7 & 9.8 (Results provided by Sébastien Wolf, Debrégeas research group). Above a given threshold, the neuron will have the value 1, otherwise 0.

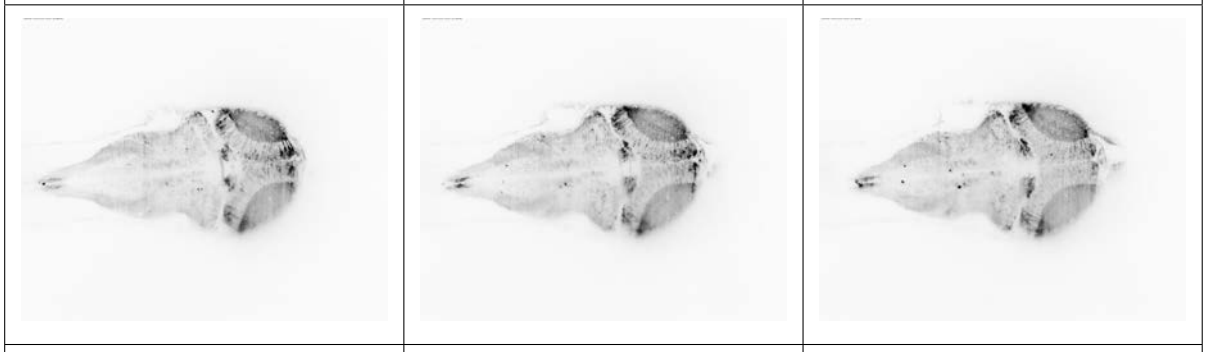
9.2 Neural network reconstructions with the 3off2 method

9.2.1 Preprocessing of SPIM images by neuron clustering

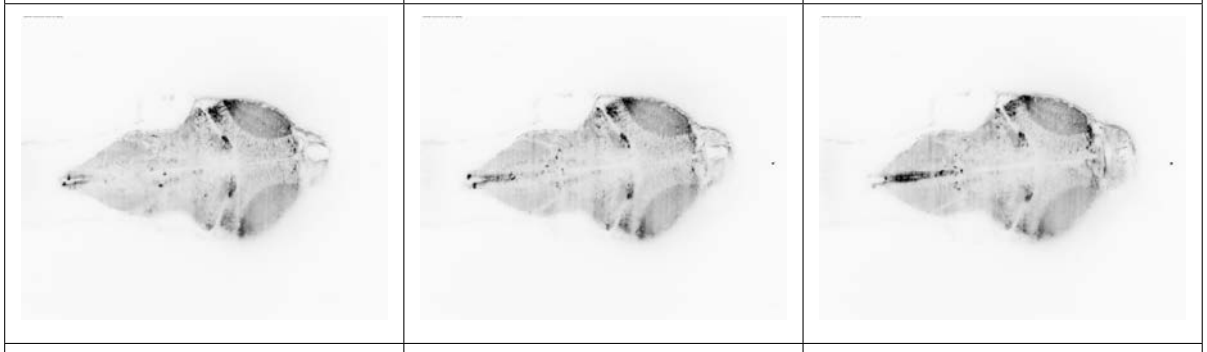
As introduced in section 9.1.4, the SPIM imaging technology allows for the acquisition of a large number of neurons spanning different areas of the brain. The simultaneous recording of distinct brain regions provides the mean to measure correlated activities throughout the brain. The experimental setup described in Fig 9.2 was used to acquire 12 distinct dorso-ventral z-planes of a zebrafish brain in the larval stage. The z-planes correspond to 2D-layers separated by a distance of 7μm, covering a total range of 77μm. An overview of the acquired images is given in table 9.1.

Groups of neurons involved in similar functionalities within a specific region typically undergo similar bursts when the electrical signals are transmitted throughout the brain. Several studies have shown that the brain of zebrafish larvae can be partitioned into distinct neural areas [Bianco and Engert, 2015, Kubo et al., 2014, Portugues et al., 2014]. Thus, in order to avoid redundant edges corresponding to correlations within individual functional module, we have clustered the neurons that display correlated bursts within the same neighbourhood using a hierarchical

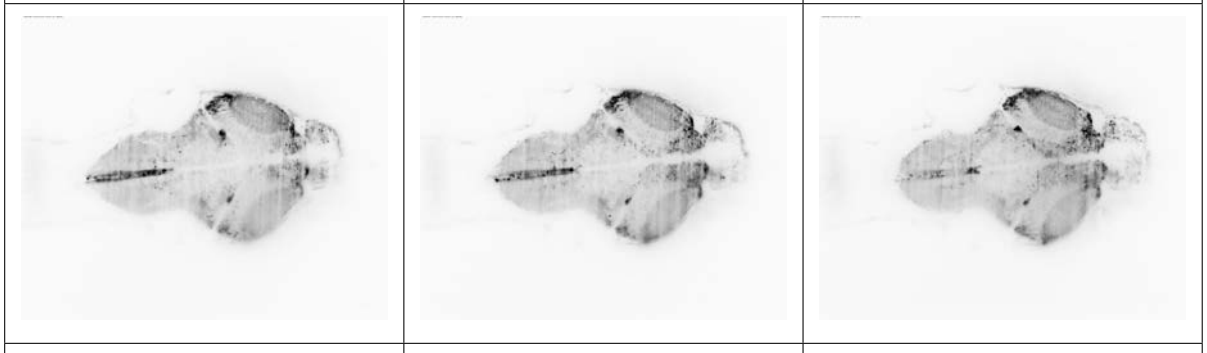
(1) **Dorsal view** $z = z_0$ | (2) $z = z_0 + 7\mu m$ | (3) $z = z_0 + 14\mu m$



(4) $z = z_0 + 21\mu m$ | (5) $z = z_0 + 28\mu m$ | (6) $z = z_0 + 35\mu m$



(7) $z = z_0 + 42\mu m$ | (8) $z = z_0 + 49\mu m$ | (9) $z = z_0 + 56\mu m$



(10) $z = z_0 + 63\mu m$ | (11) $z = z_0 + 70\mu m$ | (12) **Ventral View** $z = z_0 + 77\mu m$

TABLE 9.1 **Multi-view z-planes of a zebrafish brain in larval stage.** The sections are separated by a distance of 7μm. Z-plane were acquired at 20Hz (100ms exposure time).

clustering. Tables 9.2 & 9.3 give an overview of the clustering results over the images collected at each z-plane, when considering 15 (Table 9.2) or 30 (Table 9.3) clusters.

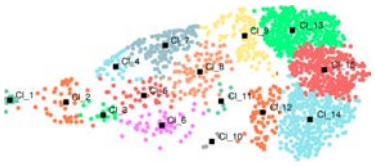
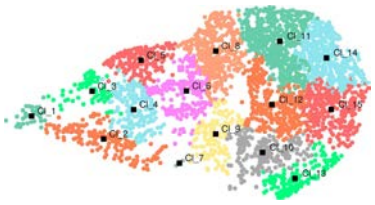
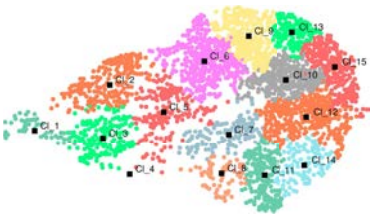
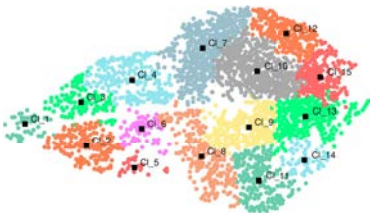
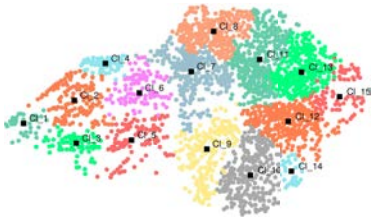
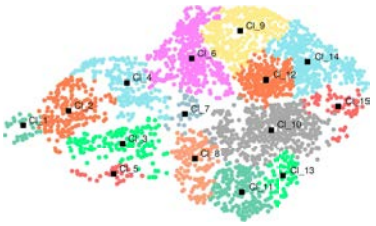
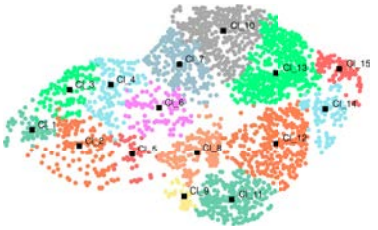
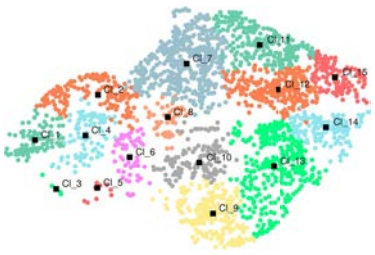
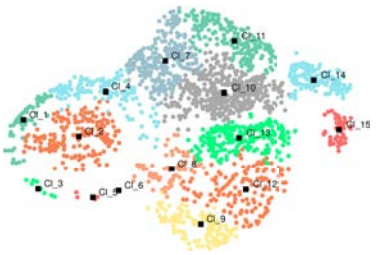
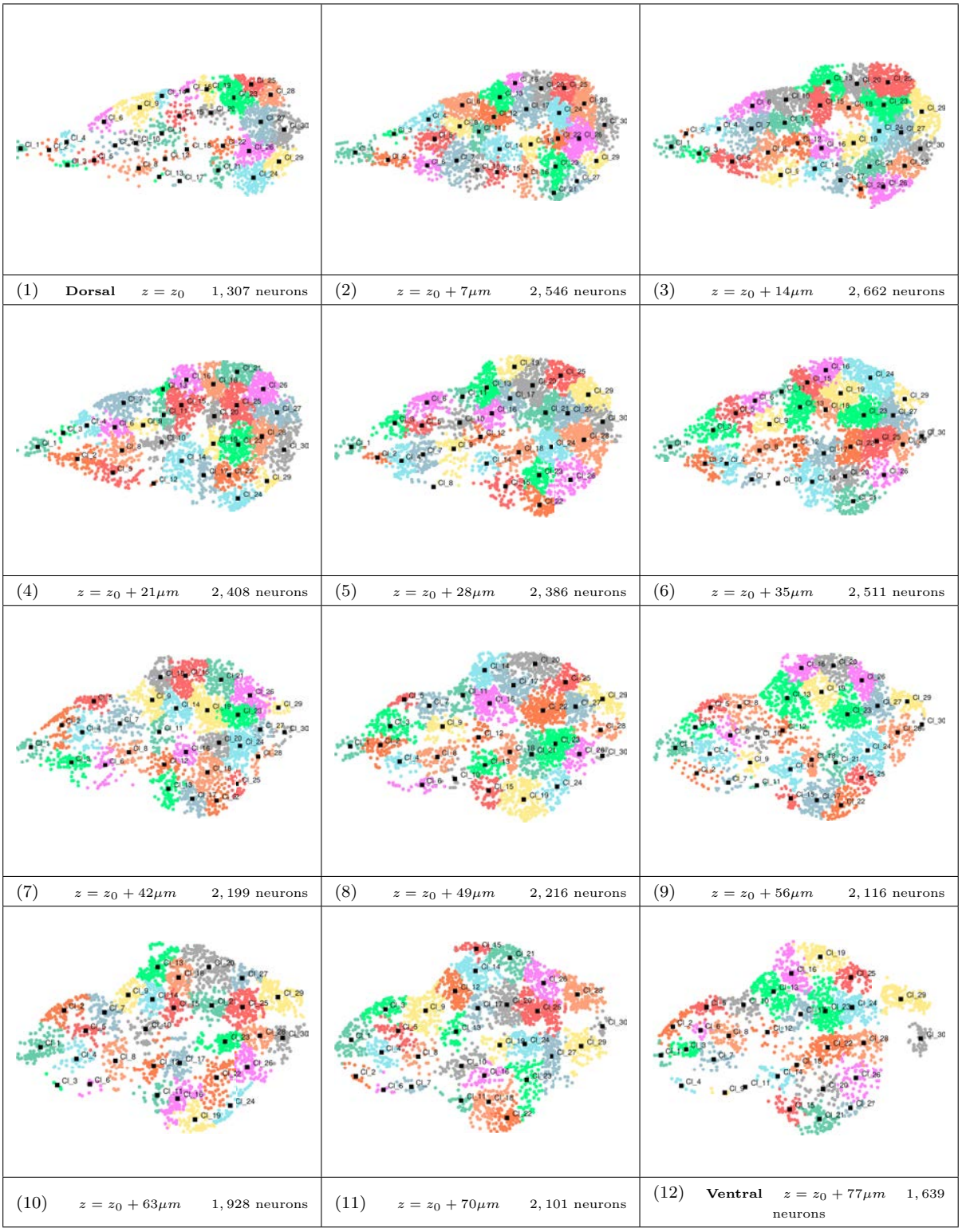
		
(1) **Dorsal** $z = z_0$ 1,307 neurons	(2) $z = z_0 + 7\mu m$ 2,546 neurons	(3) $z = z_0 + 14\mu m$ 2,662 neurons
		
(4) $z = z_0 + 21\mu m$ 2,408 neurons	(5) $z = z_0 + 28\mu m$ 2,386 neurons	(6) $z = z_0 + 35\mu m$ 2,511 neurons
		
(7) $z = z_0 + 42\mu m$ 2,199 neurons	(8) $z = z_0 + 49\mu m$ 2,216 neurons	(9) $z = z_0 + 56\mu m$ 2,116 neurons
		
(10) $z = z_0 + 63\mu m$ 1,928 neurons	(11) $z = z_0 + 70\mu m$ 2,101 neurons	(12) **Ventral** $z = z_0 + 77\mu m$ 1,639 neurons

TABLE 9.2 **Neurons clustering within z-planes of a zebrafish brain in larval stage.** The neurons of each z-plane have been clustered based on their bursting correlation and a hierarchical clustering procedure into 15 clusters. Refer to table 9.1 for raw images.



(1) **Dorsal** $z = z_0$ 1,307 neurons	(2) $z = z_0 + 7\mu m$ 2,546 neurons	(3) $z = z_0 + 14\mu m$ 2,662 neurons
(4) $z = z_0 + 21\mu m$ 2,408 neurons	(5) $z = z_0 + 28\mu m$ 2,386 neurons	(6) $z = z_0 + 35\mu m$ 2,511 neurons
(7) $z = z_0 + 42\mu m$ 2,199 neurons	(8) $z = z_0 + 49\mu m$ 2,216 neurons	(9) $z = z_0 + 56\mu m$ 2,116 neurons
(10) $z = z_0 + 63\mu m$ 1,928 neurons	(11) $z = z_0 + 70\mu m$ 2,101 neurons	(12) **Ventral** $z = z_0 + 77\mu m$ 1,639 neurons

TABLE 9.3 **Neurons clustering within z-planes of a zebrafish brain in larval stage.** The neurons of each z-plane have been clustered based on their bursting correlation and a hierarchical clustering procedure into 30 clusters. Refer to table 9.1 for raw images.

9.2.2 Reconstruction of temporal activity patterns

In previous studies, neural circuits involved in innate reflexes [Kubo et al., 2014, Portugues et al., 2014] or hunting behaviour [Bianco and Engert, 2015] have been

studied in the zebrafish larva, relying on two-photon technology. In this section, the network reconstructions of 3off2 are compared with the temporal patterns uncovered in these experiments. These results are still exploratory and the offsets of various parameters remain to be studied.

9.2.2.1 Alternating neural activity in optokinetic reponse

The optokinetic response (OKR) is an innate reflexive behaviour that tends to reduce the motion of the image on the retina with an horizontal rotational movement of the eyes. In [Portugues et al., 2014], the OKR has been induced in zebrafish larva using sinusoidally rotating whole-field stimuli around the animal. The neural responses of the larval zebrafish, genetically modified to encode a calcium indicator GCaMP5G, have been recorded with a two-photon imaging system. The 2D layers of the brain analysed in [Portugues et al., 2014] revealed alternating temporal patterns. Specifically, (a) a columnar activity has been observed in the dorsal medulla, with an antiphase bursting pattern between left and right sides, and (b) left/right-alternating activity in the neuropil has also been observed (Fig. 9.9). Besides, (c) sensory areas, such as tectal neuropil, pretectum and inferior olive have been reported to typically precede the activity in motor structures in the hindbrain.

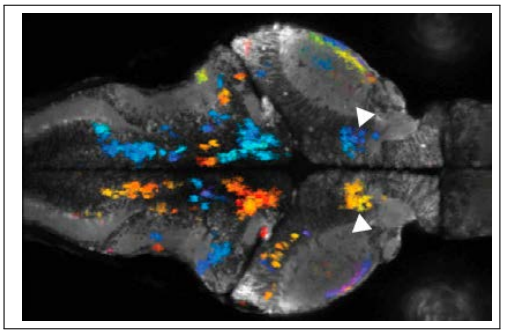


FIGURE 9.9 **Active areas during OKR behaviour.** Color-coded activity phase of regions segmented from volumetric imaging data across three fish. The different colors highlight the temporal alternating patterns. The medulla (caudal part of the brain) shows a columnar alternating activity. Two areas in the pretectum (arrowheads) also display an alternating pattern. (Image taken from [Portugues et al., 2014])

The causal networks reconstructed with the 3off2 method report these alternating patterns. Fig. 9.10 & 9.11 show the reconstructed networks from the binary data corresponding to the neural activity in layer 5, when considering 15 (Fig. 9.10) and 30 (Fig. 9.11) clusters. The discovered edges are all associated with activation, and the color gives the strength of the relationship, from

light yellow to red, red being the strongest. Grey edges are the non-oriented relationships. In Fig. 9.11, the clusters within the medulla area are related by a directed path displaying a possible left/right temporally alternating pattern, $(Cl_{10} \rightarrow Cl_5 \rightarrow Cl_7 \rightarrow Cl_3 \rightarrow Cl_4 \rightarrow Cl_2)$. In the reconstructed network based on 15 clusters, this temporal pattern can not be seen, except within the oriented edges $(Cl_2 \rightarrow Cl_3 \rightarrow Cl_1)$ (Fig. 9.10).

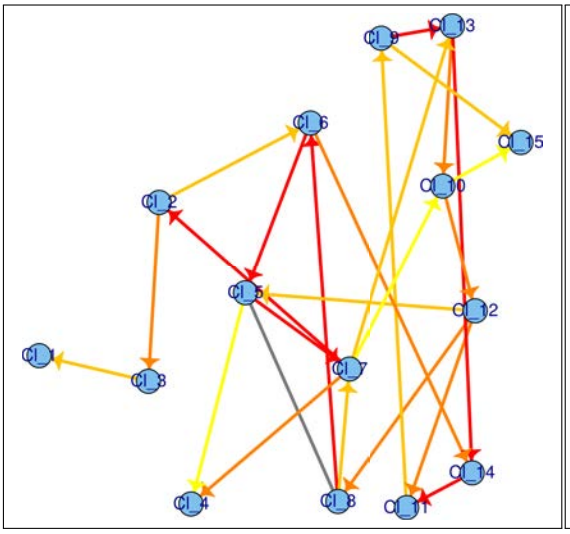


FIGURE 9.10 **3off2 reconstructed network** [Layer 5, $z = z_0 + 28\mu m$, 15 clusters].

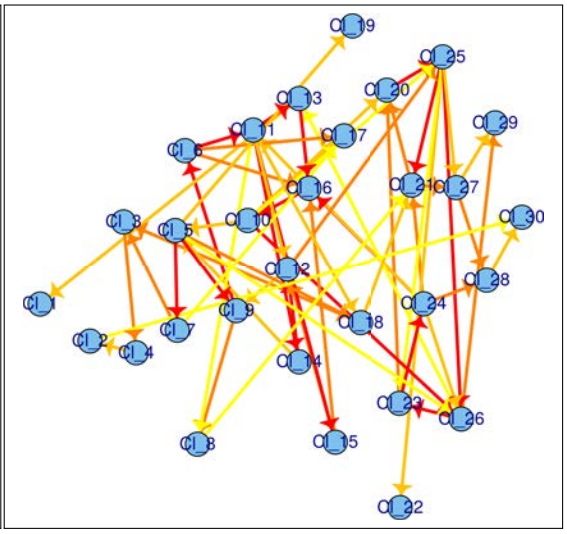


FIGURE 9.11 **3off2 reconstructed network** [Layer 5, $z = z_0 + 28\mu m$, 30 clusters].

Similar results can be seen in the 3off2 reconstructed network from the layer 6. Fig. 9.12 & 9.13 show the reconstructed network from the binary data corresponding to the neural activity in layer 6, when considering 15 (Fig. 9.12) and 30 (Fig. 9.13) clusters.

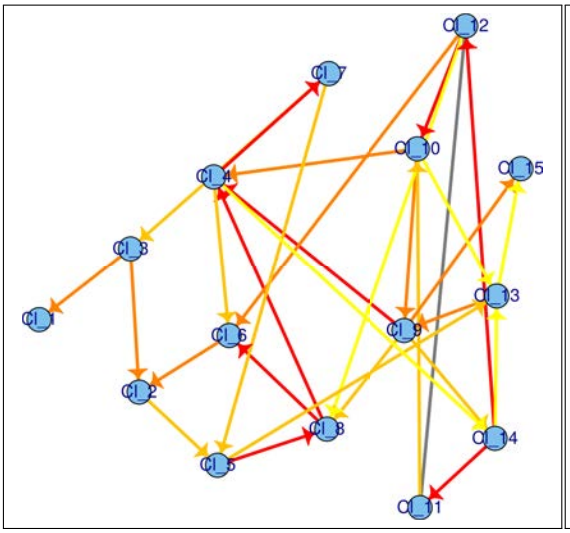


FIGURE 9.12 **3off2 reconstructed network** [Layer 6, $z = z_0 + 28\mu m$, 15 clusters].

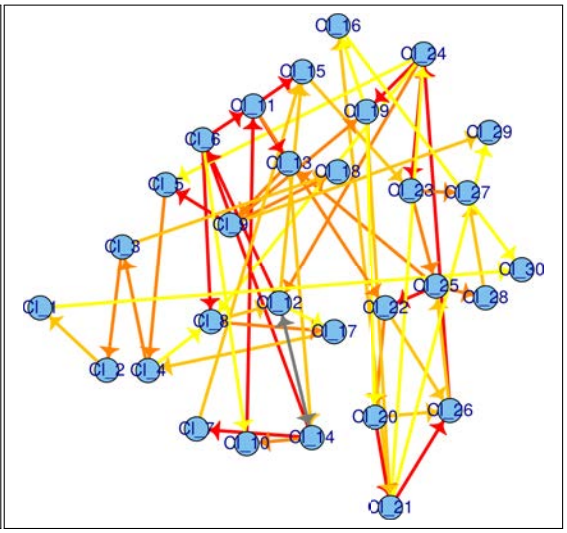


FIGURE 9.13 **3off2 reconstructed network** [Layer 6, $z = z_0 + 28\mu m$, 30 clusters].

In Fig. 9.13, the left/right alternating pattern in the medulla can be seen in the directed path ($Cl_9 \rightarrow Cl_5 \rightarrow Cl_4 \rightarrow Cl_3 \rightarrow Cl_2 \rightarrow Cl_1$). Similarly as for layer 5, the same alternating pattern cannot be inferred when the reconstruction is based on 15 clusters, except with the oriented edge $Cl_3 \rightarrow Cl_2$.

Specific temporal patterns, from sensory to motor structures have also been reported in [Portugues et al., 2014]. The reconstruction of the layer 11 by the 3off2 method, using 15 and 30 clusters, shows relationships between clusters related to the inferior olive and caudal hindbrain (Fig. 9.14 & 9.15). However, while the reconstruction based on 15 clusters displays the expected temporal pattern, *i.e.* from sensory (inferior olive) to motor (caudal hindbrain) structures (edges $Cl_1 \rightarrow Cl_3$ & $Cl_2 \rightarrow Cl_4$), the reconstruction relying on 30 clusters display an opposite (weak) temporal pattern (edges $Cl_3 \rightarrow Cl_1$ & $Cl_7 \rightarrow Cl_2$). This could be explained by the split of the bursting information between neighbour areas when using too many clusters, which in turn produces spurious (weak) orientations.

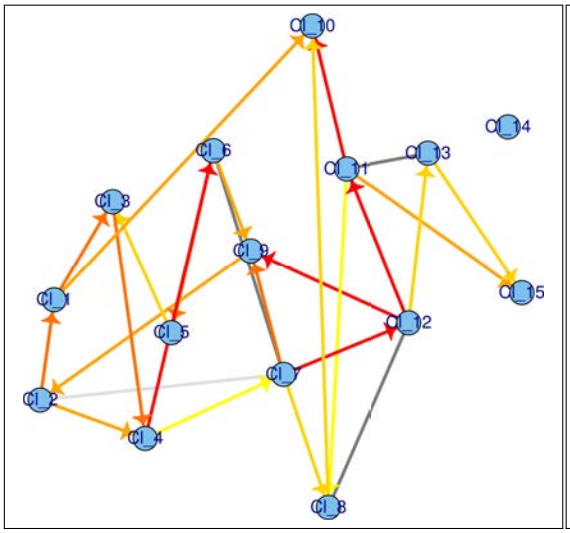


FIGURE 9.14 **3off2 reconstructed network** [Layer 11, $z = z_0 + 70\mu m$, 15 clusters].

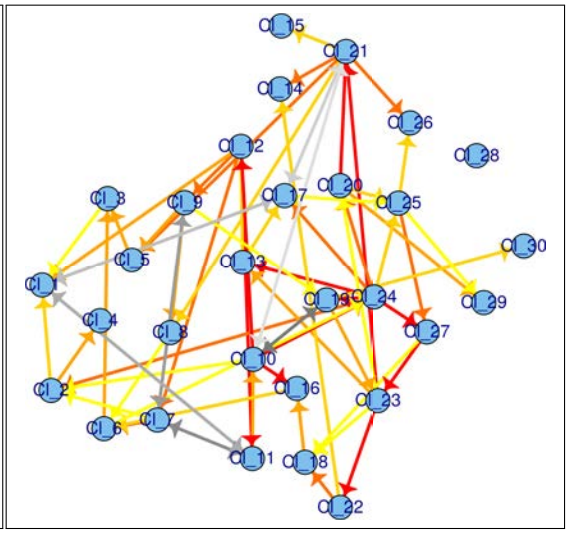


FIGURE 9.15 **3off2 reconstructed network** [Layer 11, $z = z_0 + 70\mu m$, 30 clusters].

It should be noted that, when reaching much deeper layers, 3off2 identified more relationships resulting from a common cause (bi-directed gray edges, Fig. 9.15), possibly indicated the influence of upper neurons not included in the 2D reconstructed network.

[Portugues et al., 2014] also reported a left/right alternating activity in neuropil in the vicinity of the interpeduncular nucleus. Interestingly, the networks reconstructed with the 3off2 method display directed cycles between the areas closely related to the neuropil. Specifically, Fig. 9.10 displays the directed cycles ($Cl_9 \rightarrow$

$Cl_{13} \rightarrow Cl_{10} \rightarrow Cl_{12} \rightarrow Cl_{11} \rightarrow Cl_9$) and ($Cl_9 \rightarrow Cl_{13} \rightarrow Cl_{14} \rightarrow Cl_{11} \rightarrow Cl_9$), where Cl_9 & Cl_{11} belong to the optic tectum, Cl_{10} & Cl_{12} belong to the pretectum and Cl_{13} & Cl_{14} correspond to retinal ganglion cells. Similarly, Fig. 9.12 display the directed cycle ($Cl_9 \rightarrow Cl_{14} \rightarrow Cl_{12} \rightarrow Cl_{10} \rightarrow Cl_9$).

Although the reconstructed networks relying on 30 clusters do not show directed cycles between these areas, they display orientations that could be interpreted as alternating patterns. Indeed, the reconstructed layer 5 in Fig. 9.11 links the clusters Cl_{21} & Cl_{24} belonging to the neuropil, and the clusters Cl_{27} & Cl_{28} belonging to the pretectum through the neural circuit ($Cl_{21} \rightarrow Cl_{27} \rightarrow Cl_{28} \leftarrow Cl_{24} \rightarrow Cl_{21}$). Similarly, the reconstructed layer 6 in Fig. 9.13 links the clusters Cl_{23} & Cl_{25} belonging to the neuropil, and the clusters Cl_{27} & Cl_{28} belonging to the pretectum through the neural circuit ($Cl_{23} \rightarrow Cl_{27} \leftarrow Cl_{28} \leftarrow Cl_{25} \leftarrow Cl_{23}$). Interestingly, Fig. 9.11 & 9.13 both show directed edges from the retinal ganglion cells to the neuropil area ($Cl_{25} \rightarrow Cl_{21}$ & $Cl_{26} \rightarrow Cl_{23} \rightarrow Cl_{24}$, Fig. 9.11; $Cl_{24} \rightarrow Cl_{23}$ & $Cl_{26} \rightarrow Cl_{25}$, Fig. 9.13), as reported in the experimental observations [Portugues et al., 2014].

Hence, the 3off2 reconstruction method could display similar alternating temporal pattern as reported in [Portugues et al., 2014]. Yet, while the resulting patterns in [Portugues et al., 2014] have been uncovered with temporal data, it should be emphasized that the 3off2 method does not require temporal information to infer the direction of interaction from observational data.

9.2.2.2 Neural activity related to hunting behaviour

During prey catching, the zebrafish undergoes a specific behaviour involving eye convergence with saccades that increase when approaching the target. This specialized oculomotor behaviour is suspected to favour the distance estimation to the prey and has been shown to imply the recruitment of assemblies of neurons in the optic tectum (OTc). Using two photon calcium imaging, [Bianco and Engert, 2015] showed that the bursts produced by these assemblies precede and control the triggering of hunting responses and propose a neural model circuit underlying the hunting behaviour of the zebrafish (Fig. 9.16, taken from [Bianco and Engert, 2015]).

Specifically, the visual information is transmitted to the retinal ganglion cell (see [2], Fig. 9.16) of the contralateral hemisphere. Neurons characterizing the stimulus are activated in the rostral tectum and recruit assemblies of tectal neurons ([3] & [4]

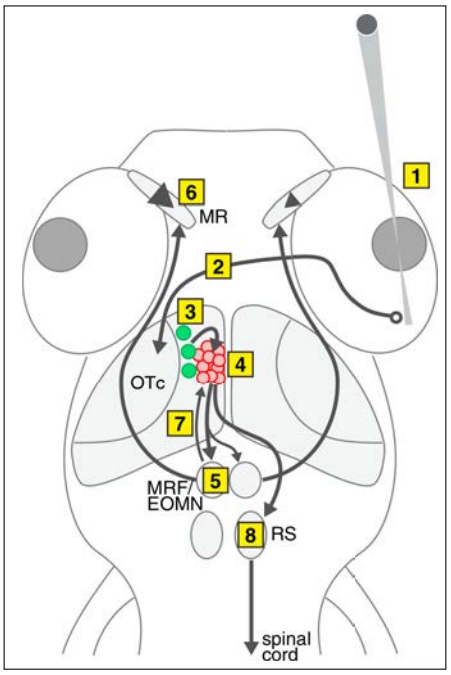


FIGURE 9.16 **Model circuit for visual prey recognition and release of hunting reponses.** MR, medial rectus; OTc, optic tectum; MRF, mesencephalic reticular formation; EOMN, extra-ocular medial rectus motoneurons; RS, reticulospinal. This scheme provides a model circuit emerging from a prey-like vosual stimulus when observed by the zebrafish from the right temporal retina. See main text for details.

resp., Fig. 9.16). The resulting bursts activate the downstream extra-ocular medial rectus motoneurons (EOMNs) in the oculomotor nucleus ([5], Fig. 9.16) which in turn control the medial rectus (MR) to produce a convergent saccade ([6], Fig. 9.16). The tectal assemblies also participate in the activation of the reticulosponal (RS) neurons that control spinal circuits, and thus the tail participating in the orientation toward the prey ([8], Fig. 9.16). The projections from MRF to the upstream tectum ([7] resp., Fig. 9.16) could be either a feedback control of the eye movement or an efferent copy mechanism that has been shown to favour a stable perception of the prey during the zebrafish movements.

The causal networks reconstructed with the 3off2 method partially recover this model circuit. Figs. 9.17 & 9.18 display the reconstructed network based on 30 clusters of the layers 2 and 4. In particular, the directed paths $(Cl_9 \to Cl_{25} \to Cl_{30})$ & $(Cl_{21} \to Cl_{30})$ in Fig. 9.17 and $(Cl_9 \to Cl_{21} \to Cl_{26})$ in Fig. 9.18 possibly correspond to efferent projections from MRF to MR. Besides, the characterization of the prey stimulus followed by the activation of MRF assemblies follow the directed paths $(Cl_{26} \to Cl_{24})$, $(Cl_{26} \to Cl_{23})$ & $(Cl_{26} \to Cl_{22} \to Cl_{17})$ in the reconstructed network of Fig. 9.17. Finally, the Fig. 9.18 also displays two oriented

edges, $(Cl_{29} \to Cl_{20})$ & $(Cl_{26} \to Cl_{14})$, that could be attributed to the transmission of the visual information to the retinal ganglion cells.

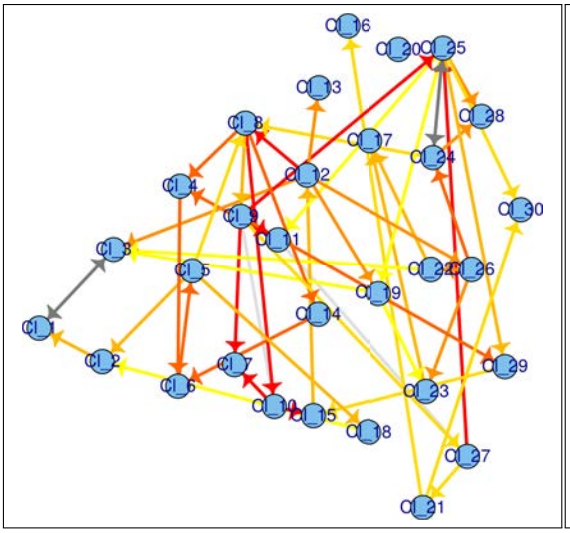


FIGURE 9.17 **3off2 reconstructed network** [Layer 2, $z = z_0 + 7\mu m$, 30 clusters].

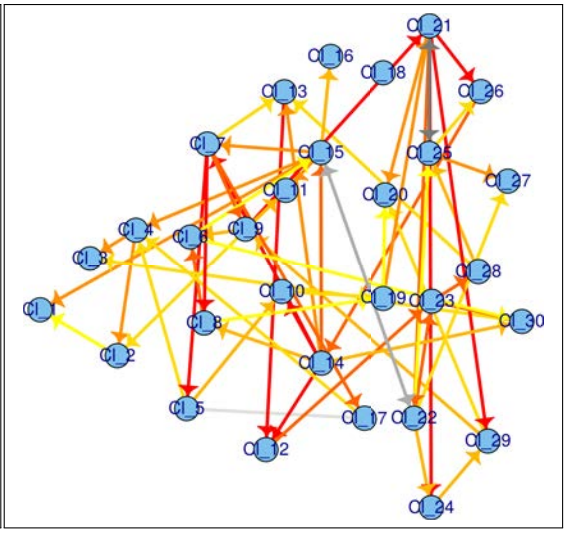


FIGURE 9.18 **3off2 reconstructed network** [Layer 4, $z = z_0 + 21\mu m$, 30 clusters].

The activation of RS neurons ([8], Fig. 9.16) can be seen in the reconstructed network of the layer 1 (Fig. 9.19). In particular, the oriented edges $(Cl_{18} \to Cl_1)$ & $(Cl_{12} \to Cl_8)$ and the directed path $(Cl_{17} \to Cl_{13} \to Cl_3)$ correspond to the recruitment of spinal circuits involved in the tail movement. In deeper layer, such as layer 9 (Fig. 9.20), more directed paths of this type can be observed in the reconstructed networks, *e.g.* $(Cl_{13} \to Cl_8 \to Cl_5 \to Cl_3)$, $(Cl_{12} \to Cl_{10} \to Cl_6)$ & $(Cl_9 \to Cl_7 \to Cl_2)$. This might be due to the fact that this depth corresponds to a position below the cerebellum, offering a broader view of the partially underneath medulla.

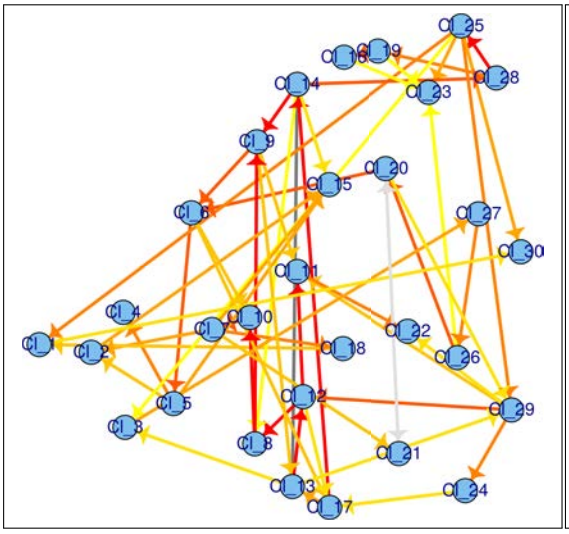


FIGURE 9.19 **3off2 reconstructed network** [Layer 1, $z = z_0$, 30 clusters].

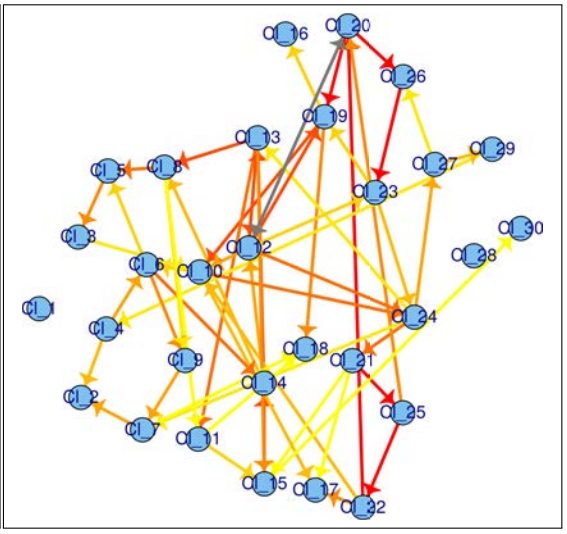


FIGURE 9.20 **3off2 reconstructed network** [Layer 9, $z = z_0 + 56\mu m$, 30 clusters].

Further investigation in terms of parameters should still be done in order to define the optimal number of clusters, to decide for the most accurate clustering method and to identify the mostly significant active neurons. Yet, these preliminary 2D network reconstructions could partially recover the model circuit deduced from experiments relying on stimuli [Bianco and Engert, 2015, Portugues et al., 2014] and the analysis of the spontaneous activity of the zebrafish brain based on the 3off2 method. This support the prevailing hypothesis of an underlying brain network revealed by spontaneous activity or stimulations. In the future, we would like to implement clustering of the signals in 3D and subsequent 3D network reconstruction. As several 2D reconstructed networks displayed in this chapter show edges suggesting the existence of latent variables (bi-directed gray edges), 3D clustering could possibly identify neurons in upper or deeper layers participating in these correlations.

Chapter 10

Reconstruction of the hematopoiesis regulation network with the 3off2 method

In this chapter, the 3off2 approach has been applied on a real biological dataset related to hematopoiesis. The hematopoiesis is a process that controls the emergence and the replacement of blood cells. In adults, these cells, originated from bone marrow, differentiate in distinct categories, among which the erythrocytes, the leukocytes and the thrombocytes, that ensure specific functions. For instance, while the erythrocytes (red blood cells) guarantee the oxygen transportation, the thrombocytes (platelets) are involved in the hemostasis and the leukocytes (white blood cells) play an important role in the immune system. During the hematopoiesis, transcription factors[1] trigger and guide the differentiation process from Hematopoietic Stem Cells (HSC) to mature cells.

This chapter presents and discusses the reconstruction of a regulatory subnetwork of the hematopoiesis performed with 3off2 reconstruction method. This part of my project has been developed in collaboration with a PhD student in bioinformatics in the Isambert's group, Louis Verny [Affeldt and Isambert, 2015, Affeldt et al., 2015].

[1] A transcription factor is a protein that controls the rate of transcription of a gene by binding to the DNA sequence.

10.1 The hematopoiesis regulation network

Transcription factors play a central role in hematopoiesis, from which derive the blood cell lineages. As suggested in previous studies, changes in the regulatory interactions among transcription factors [Oram et al., 2010] or their overexpression [Cleveland et al., 2013] might be involved in the development of T-acute lymphoblastic leukaemia (T-ALL). The key role of the hematopoiesis and the potentially serious consequences of its disregulations emphasize the need to accurately establish the complex interactions between the transcription factors involved in this critical biological process.

10.2 Regulation network reconstruction with the 3off2 method

10.2.1 Dataset and interactions of interest

The dataset we have used for this analysis consists of the single cell expressions[2] of 18 transcription factors [Moignard et al., 2013], known for their role in hematopoiesis. 597 single cells representing 5 different types of hematopoietic progenitors have been included in the analysis ($N = 597$). We reconstructed the corresponding network with the 3off2 method (Figure 10.1, Table 10.1) and other available approaches (Table 10.1). In particular, we focused on 11 regulatory interactions for which specific experimental evidence have been reported in the literature (Table 10.1).

10.2.2 The interactions recovered by 3off2 and alternative methods

The corresponding network has been reconstructed with the 3off2 method and four other available approaches, namely, PC [Spirtes and Glymour, 1991] implemented in the `pcalg` package [Kalisch and Bühlmann, 2008, Kalisch et al., 2012], Bayesian inference using hill-climbing heuristics as well as the Max-Min Hill-Climbing (MMHC) hybrid method [Tsamardinos et al., 2006], both implemented in the `bnlearn` package [Scutari, 2010], and finally, Aracne [Margolin et al., 2006] implemented in the `minet` package [Meyer et al., 2008] (Tables 10.1 and 10.2).

[2] microfluidics qRT-PCR CT measures

3off2 uncovers all 11 interactions for which specific experimental evidence has been reported in the literature (Figure 10.1, red links: known activations; blue links: known repressions) as well as 30 additional links (Figure 10.1, grey links: unknown regulatory interactions). By contrast, randomization of the actual data across samples for each TF leads to only 5.25 spurious interactions on average between the 18 TFs, instead of the 41 inferred edges from the actual data, and 1.62 spurious interactions on average, instead of the 16 interactions predicted among the 10 TFs involved in known regulatory interactions, Figure 10.1. This suggests that around 10-30% of the predicted edges might be spurious, due to inevitable sampling noise in the finite dataset.

In particular, the 3off2 inference approach successfully recovers the relationships of the regulatory triad between *Gata2*, *Gfi1b* and *Gfi1* as described in [Moignard et al., 2013] and reports correct orientations for the edges involving *Gata2* (*Gfi1b* and *Gfi1* crossregulate in fact one another [Moignard et al., 2013], Table 10.1).

11 known Regulatory interactions		3off2	PC	PC	MMHC	MMHC	Bayes hc	Bayes hc	Aracne
		NML	$\alpha=10^{-1}$	$\alpha=10^{-2}$	*BDe*	*BIC*	*BDe*	*BIC*	$\epsilon=0$
Gata2 → Gfi1b	[Moignard et al., 2013]	→	↛	—	≠	≠	→	≠	≠
Gfi1 → Gata2	[Moignard et al., 2013]	→	→	—	→	↛	→	↛	—
Gfi1b ⇆ Gfi1	[Moignard et al., 2013]	↛	↛	—	↛	↛	↛	↛	—
Gfi1 → PU.1	[Spooner et al., 2009]	→	→	≠	≠	≠	→	→	—
Lyl1 → Gfi1	[Zohren et al., 2012]	→	↛	≠	≠	≠	→	↛	—
Ldb1 → Meis1	[Li et al., 2011]	→	≠	≠	≠	≠	↛	≠	≠
Gata2 → Scl	[Göttgens et al., 2002]	↛	→	—	→	→	→	→	—
Gfi1b → Meis1	[Chowdhury et al., 2013]	↛	↛	—	→	→	→	→	—
Ldb1 → Lyl1	[Li et al., 2011]	→	≠	≠	≠	≠	≠	≠	≠
Erg → Lyl1	[Chan et al., 2006]	→	↛	—	→	→	→	↛	—
Gata1 → Gata2	[Grass et al., 2003]	↛	↛	—	→	→	→	→	—
Correct edges (out of **11**) (→/↛/—)		**11**	**9**	**7**	**6**	**6**	**10**	**8**	**8**
- Correct orientations	(→)	7	3	0	5	4	8	4	0
- Mis/non-orientations	(↛/ —)	4	6	7	1	2	2	4	8
Missing links	(≠)	**0**	**2**	**4**	**5**	**5**	**1**	**3**	**3**

TABLE 10.1 **Interactions reconstructed by** 3off2 **and alternative methods for a subnetwork of hematopoiesis regulation.** → indicates a successfully recovered interaction including its direction as reported in the literature (see references in the table). ↛ corresponds to a successfully recovered interaction, however, with an opposite direction as reported in the literature. ≠ stipulates that no direct regulatory interaction has been inferred, while — corresponds to an undirected link. Red/blue edges correspond to experimentally proven activating/repressing interactions. Note in particular that Aracne does not infer edge direction.

The network reconstructed by 3off2 also correctly infers the influence of *Gfi1* on *PU.1* [Spooner et al., 2009], the regulation of *Gfi1* by *Lyl1* [Zohren et al., 2012], and the regulatory effects of *Ldb1* on *Meis1* and *Lyl1* [Li et al., 2011]. Finally, the

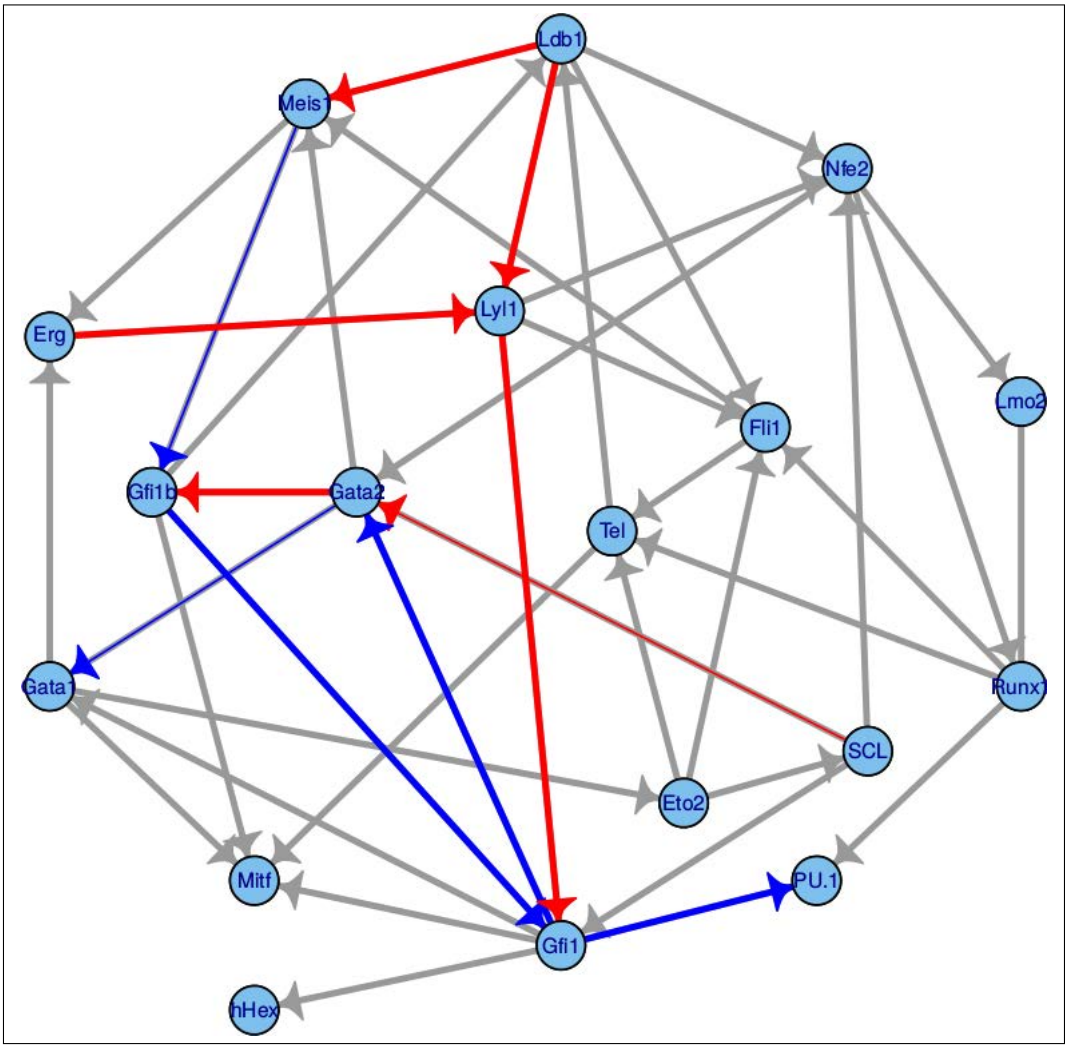


FIGURE 10.1 **Hematopoietic subnetwork reconstructed by 3off2** The dataset [Moignard et al., 2013] concerns 18 transcription factors, 597 single cells, 5 different hematopoietic progenitor types. Red and blue edges correspond to experimentally proven activations and repressions, respectively as reported in the literature (Table 10.1), while grey links indicate regulatory interactions for which no clear evidence has been established so far. Thinner arrows underline 3off2 misorientations.

interactions (*Gata1* − *Gata2*), (*Erg* − *Lyl1*), (*Gata2* − *SCL*) and (*Meis1*−*Gfi1b*) are correctly inferred, however, with opposite directions as reported in the literature. Yet, overall 3off2 outperforms most of the other methods tested for the reconstruction of hematopoietic regulatory networks (Tables 10.1 and 10.2). Only the Bayesian hill-climbing method using a BDe score leads to comparable results by retrieving 10 out of 11 interactions and correctly orienting 8 of them.

These encouraging results from the 3off2 reconstruction method on experimentally proven regulatory interactions (red/blue edges in Figure 10.1) could motivate further investigations on novel regulatory interactions awaiting to be tested for their possible role in hematopoeisis (*e.g.* grey edges in Figure 10.1).

Statistical Measure	3off2 *NML*	PC $\alpha=10^{-1}$	PC $\alpha=10^{-2}$	MMHC *BDe*	MMHC *BIC*	Bayes hc *BDe*	Bayes hc *BIC*	Aracne $\epsilon=0$
Recall^u	1	0.82	0.64	0.55	0.55	0.91	0.73	0.73
Precision^u	0.65	0.64	0.7	0.55	0.55	0.56	0.67	0.89
Fscore^u	0.79	0.72	0.67	0.55	0.55	0.69	0.70	0.80
H^u	0.13	0.18	0.16	0.22	0.22	0.22	0.13	0.09
HIM^u	0.16	0.17	0.13	0.19	0.19	0.22	0.10	0.08
Distance^u	0.13	0.16	0.16	0.22	0.22	0.18	0.16	0.11
Recall^d	1	0.6	0.2	0.5	0.44	0.89	0.57	NA
Precision^d	0.41	0.21	0.1	0.45	0.36	0.47	0.33	NA
Fscore^d	0.58	0.31	0.13	0.47	0.4	0.62	0.42	NA
H^d	0.14	0.21	0.17	0.12	0.14	0.14	0.14	NA
HIM ^{d}	0.13	0.16	0.15	0.09	0.11	0.14	0.11	NA
Distance^d	0.22	0.29	0.29	0.24	0.27	0.22	0.24	NA

TABLE 10.2 **Interactions reconstructed by 3off2 and four alternative methods on the subnetwork of 11 known regulatory interactions in hematopoiesis (Figure 10.1).** These calculations were made assuming that the 11 experimentally proven interactions correspond to the true regulatory network (in practice, we expect that some of the inferred edges counted as false positives might in fact turn out to be correctly predicted links). ^{u} indicates that the calculations were made on the undirected network, whereas ^{d} means directed network. The "Distance" measure corresponds to the number of "errors" (including orientations for the directed network comparison) over the total number of possible edges, *i.e.* Distance=$2(FP+FN)/n(n-1)$, where $n=10$ is the number of nodes in the subnetwork including the 11 known regulatory interactions (Figure 10.1).

Chapter 11

Reconstruction of mutational pathways involved in tumours progression with the 3off2 method

Many biomolecular networks sustain the life processes of an organism. Mutations of genes belonging to these functional networks may lead to complex diseases, such as cancer. An increasing number of studies are trying to identify genes susceptible to mutations that favour the development of genetic diseases [Dees et al., 2012, Mermel et al., 2011, V., 2014]. In particular, genes involved at upstream positions in these biomolecular networks are of great interest, as they may be interesting therapeutic drug targets. However, there exists a significant heterogeneity in the mutations between individuals developing the same type of cancer, and this heterogeneity can also be observed in the tumour of the same patient [Lawrence, 2013]. In fact, the deregulations that drive tumour growth may originate from distinct mutated signalling pathways or arise through mutations at different levels of the same signalling pathway.

Despite this intra- and inter-patients heterogeneity, there still exist typical *fences* that tumoral cells should overcome to initiate and sustain an uncontrolled proliferation [Vogelstein et al., 2013]. For instance, it is 'beneficial' for the (future) tumour cells to first lose their ability to differentiate, as this enhances their division rate and thus their proliferation. Then, mutations preventing DNA error checks and/or repair of these errors also favour the tumour development by offering new potentially 'advantageous' mutations. Subsequently, any mutation sustaining a significant angiogenesis and providing anti-growth factor resistance ensures

further development of the malignant tissue. Finally, the metastasis stage can be reached by acquiring mutations that enable cells to escape from the primary tumour tissue.

In this chapter, I present results related to the reconstruction of such *mutational pathways*, *i.e.* cascades of mutations that trigger and/or sustain the tumour development, in breast cancer tumours. These results rely on data made available by the COSMIC (Catalog Of Somatic Mutations In Cancer) database [Forbes et al., 2008], a web resource developed and maintained by the Wellcome Trust Sanger Institute. This part of my project has been developed in collaboration with a Master 2 student in bioinformatics in the Isambert's group, Wei Wenjun.

11.1 The breast cancer dataset

11.1.1 A brief overview of breast cancer

The breast cancer is the most common cancer in women. The malignant tissue typically develops from *somatic*[1] mutations, although inherited mutations may favour the emergence of breast cancer tumours.

Previous studies have confirmed the existence of four main breast cancer molecular classes, namely the luminal A, the luminal B, the triple negative (or basal-like) and the HER2 types [TCGA, 2012]. The luminal A and B are the most common subtypes (about 40% and 20% of the cases) and are associated with a deregulation of the oestrogen and progesterone receptors. Thus, they can be treated with an hormonal therapy. As compared to the luminal A type, the luminal B has a poorer prognosis. The luminal A tumour cells are characterized by a slow growth, while the luminal B subtype is more aggressive, proliferate faster, and is prone to spread through blood vessels.

The HER2 and the triple negative classes are less common (about 15% and 20% of the cases). The HER2 subtype is prone to early and frequent metastasis. This subtype typically corresponds to an over-expression of the HER2 gene, resulting in an over-production of growth-enhancing proteins, and is usually associated with a poor prognosis. The triple negative class produces fast-growing aggressive tumours. As this subtype lacks oestrogen and progesterone receptors, patients cannot benefit from an endocrine therapy.

[1] Somatic mutations are the mutations acquired during the life of an individual. By contrast with *germline* mutations, they are not inherited from the parents.

11.1.2 The issue of missing values

The COSMIC database[2] contains information on tumour samples associated with many cancers. These samples have been collected and curated from the scientific literature and large-scale experimental screens from the Cancer Genome Project at the Sanger Institute. The version 69 of COSMIC, on which rely our dataset, provides 999, 872 samples that contain 1, 808, 915 non-synonymous mutations[3] for 27, 829 genes. These data have been obtained from 9, 424 genomes and 19, 031 publications.

The growing number of studies in cancer, and the advances in next-generation sequencing (NGS) technologies enabled regular updates of the COSMIC resources since 2004. In particular, the number of samples for which the whole-genome or whole-exome [4] has been sequenced is rapidly increasing (+33% between the version 69 and 70, table 11.1). However these whole-genome/exome data still represent a small portion of the COSMIC database (about 1%), as most studies have relied so far on the sequencing of a few genes of interest, *i.e.* frequently mutated in a specific cancer.

	V69	**V70**	**V70-V69**	**%**
Genes	27,829	28,735	906	+3%
Samples	999,872	**1,029,547**	29,675	+3%
Coding Mutations	1,808,915	2,002,811	193,896	+10%
Papers	19,031	19,703	672	+3%
Whole Genome/Exome	9,424	**12,542**	3,118	**+33**%

TABLE 11.1 **Evolution of tumour sample resources in the COSMIC database.**

A single study is typically associated with a few hundred patients and thus provides a few hundred samples. The sequencing of these samples can either be *whole-genome/exome* or restricted to specific portions of the DNA. Yet, a few hundred samples, even when issued from whole-genome sequencing, do not provide enough information to allow a robust reconstruction of mutational pathways as will be shown below. This can be overcome by gathering distinct studies, although it usually implies to rely on an incomplete dataset. As detailed in section 6.4.3,

[2] http://cancer.sanger.ac.uk/cosmic

[3] A non-synonymous mutation is a nucleotide mutation that modifies the amino acid sequence of the corresponding protein.

[4] Whole-exome sequencing methods concern the subset of DNA that encodes proteins, also named exons.

the 3off2 reconstruction method can learn networks from datasets with missing data. We thus combined tumour samples obtained from different studies on breast cancer available in the COSMIC database and could benefit from a larger number of samples for our analysis.

The data used in this chapter contains 26,494 samples that were downloaded from the COSMIC database using the web service COSMICMart[5]. For each sample, the status of each gene is either 'mutated', 'non-mutated' or 'not sequenced', the latter corresponding to a missing value. The genes were ordered by their decreasing frequency of test, *i.e.* in increasing order of their proportion of missing values. The reconstructions discussed in section 11.2 are based on the following most 23 frequently sequenced genes relative to our dataset: TP53 (32%), PIK3CA (31%), AKT1 (11%), KRAS (9%), BRAF (7%), PTEN (7%), ERBB2 (6%), RET (5%), PIK3R1 (4%), MAP2K4 (4%), ALK (4%), SMAD4 (4%), BRCA1 (4%), MET (4%), ATM (3%), FGFR1 (3%), CDH1 (3%), APC (3%), ABL1 (2%), JAK1 (2%), NOTCH1 (2%), RB1 (2%), VHL (2%).

11.2 3off2 mutational pathways in breast cancer

11.2.1 Information content of complete *vs.* incomplete datasets

Although the dataset of breast tumour samples obtained from the COSMIC database contains 26,494 samples, only 196 samples share information without missing values for the 23 considered genes. In order to evaluate the possibility of reconstructing mutational pathways from such a small dataset, the data for each gene (*i.e.* changed/unchanged protein) has been independently shuffled across all samples, and a 3off2 reconstruction has been performed on this randomized dataset. Repeating this shuffling procedure 1,000 times, for an η parameter from 1 to 10 (see section 6.5.1 for details on the η parameter) gives on average from 2.2 to 9.7 inferred edges (Table 11.2, (A)). This comparison between randomized and non-randomized datasets suggests that about half (or more) of the edges that could be learnt from this complete but small dataset may correspond to spurious relationships.

[5] http://cancer.sanger.ac.uk/biomart/martview

(A) η	*random dataset*	*non-random dataset*	*rand./ non-rand.*
1	9.7	15	65%
2	5.0	9	55%
3	3.8	8	47%
4	3.3	7	47%
5	3.0	6	50%
6	2.8	5	55%
7	2.7	4	68%
8	2.3	4	57%
9	2.2	4	55%
10	2.2	4	55%

(B) η	*random dataset*	*non-random dataset*	*rand./ non-rand.*
1	4.8	24	20%
2	2.8	20	14%
3	2.0	17	12%
4	1.6	16	10%
5	1.4	15	9%
6	1.2	14	9%
7	1.1	14	8%
8	1.0	13	8%
9	0.9	13	7%
10	0.8	12	7%

TABLE 11.2 **Evaluation of the 3off2 method when reconstructing networks from randomized and non-randomized datasets.** The dataset of (A) 196 tumour samples without missing values and (B) 26,494 tumour samples with missing values were shuffled 1,000 times, for an η parameter from 1 to 10. The column 'random' gives the average number of edges obtained when reconstructing mutational pathways with the 3off2 method from the 1,000 shuffled datasets, while the column 'non-random' provides the number of edges based on the original (non-shuffled) datasets.

By contrast, the datasets of 26,494 tumour samples containing missing values should provides more robust networks. Indeed, as given is Table 11.2, (B), the number of inferred edges when applying the 3off2 method on randomized incomplete datasets ranges from 4.8 ($\eta = 1$) to 0.8 ($\eta > 8$). This suggests lower probabilities of inferring edges that correspond to spurious associations (*e.g.* only 12% for $\eta = 3$).

11.2.2 3off2 cascades of mutations in breast cancer

The Fig. 11.1 gives the reconstructed network obtained with the 3off2 method, when setting the η parameter to 3. As this reconstruction has been done over a restricted number of genes, the network has been inferred assuming the presence of hidden variables. The color of edges indicates the likely direction of the 'advantageous' cascade of mutations for tumour progression, with warmer colors representing positively correlated mutations and cold colors representing negatively correlated mutations (red and dark blue colors highlighting a stronger association).

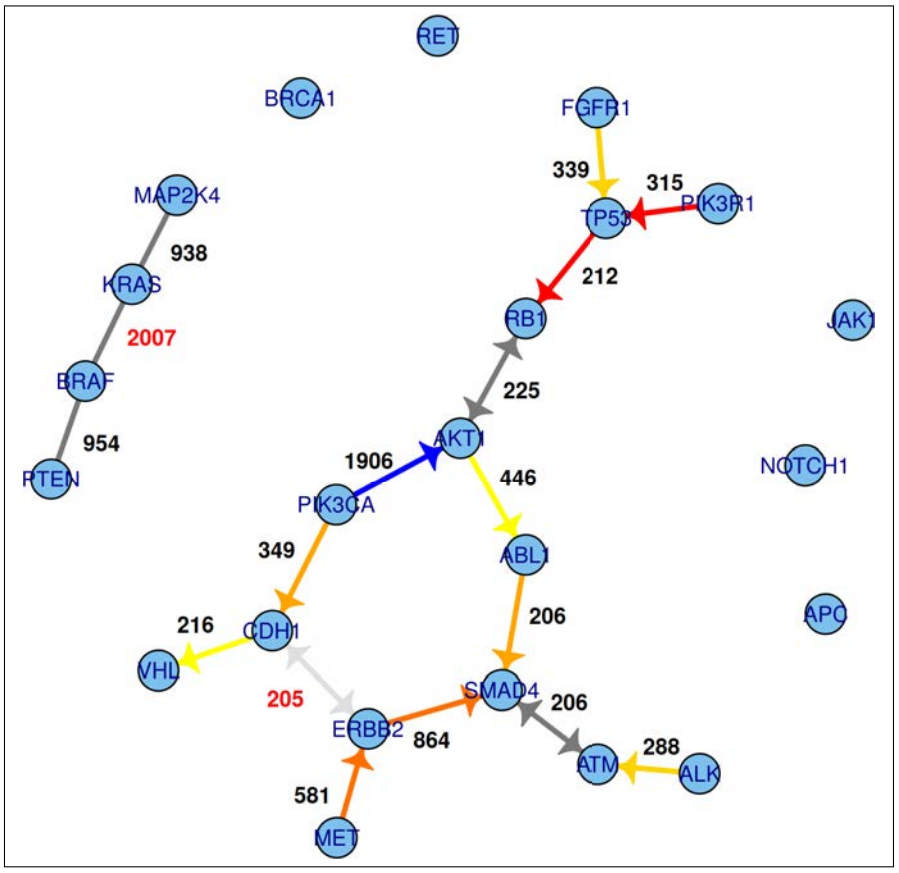


FIGURE 11.1 **Mutational pathways reconstruction with the** 3off2 **method.** The network reconstruction has been performed on 26,494 tumour samples containing information (with missing values) on 23 genes, and assuming the presence of latent variables. The oriented edges indicate the series of mutations that appears to be the most 'beneficial' for the tumour growth. The color of edges gives the direction of the 'advantageous' cascade of mutations for the tumour progression, with warmer colors representing positively correlated mutations and cold colors representing negatively correlated mutations (red and dark blue colors highlighting a stronger association). The bi-directed grey edges corresponds to associations due to one or several unmeasured latent variable(s). The count associated to each edge indicates the number of samples on which the final 2-point and 3-point information terms have been computed.

As can be seen from the Fig. 11.1, the number of samples on which the edges have been inferred ranges from 206 to 2007, with the lower bound being greater than the number of samples of the dataset without missing values. This highlights the advantage of the 3off2 method when inferring causal networks from incomplete datasets. More importantly, the network reconstructed with the 3off2 method suggest likely *temporal* information from *observational* data. Indeed, the samples correspond to 'snapshots' of the tumour mutational landscapes over distinct individuals, and no information of the cancer stage have been included in our analysis. The section 11.2.3 further details some series of mutations found in the 3off2 reconstructed network that are suspected to preferentially drive the tumour progression.

11.2.3 Temporal cascade patterns of *advantageous* mutations

Among the cascades of mutations shown in Fig. 11.1, the mutational pathway (PIK3R1 → TP53 → RB1) is an interesting example of a series of *fences* that the cells should overcome in order to initiate and sustain tumour growth. PIK3R1 is a gene that encodes one of the two subunits of the heterodimer PI3K (the other subunit being encoded by PIK3CA) that plays an important role in cellular processes, in particular cell growth and proliferation. It has been shown that alterations of PIK3R1 (or PIK3CA) result in PI3K pathway activation, and thus in an enhancement of cell proliferation [Cheung et al., 2011, Cizkova et al., 2013]. Mutations in TP53 are frequently found in breast cancer and induce the production of a non-functional protein, that can no longer ensure the control of errors in the genome or regulate the cell division process [Gasco et al., 2002]. This enables the accumulation of mutations in the proliferating cells. Finally, mutated RB1 gene is often found in the triple-negative breast cancer subtype [Robinson et al., 2013]. This gene plays an important role in the regulation of DNA replication before the division of the cell. It has been found that alterations of RB1 may induce genomic instabilities that trigger chromosomal aberrations, which might favour fixations of subsequent mutations.

Likewise, the series of mutations (PIK3CA → CDH1 → VHL) also provides an interesting cascade of mutations. As previously introduced, mutations in PIK3CA favour abnormal cell growth and proliferation. Then, mutations in CDH1, which produces E-cadherin,a major cell-cell adhesion molecule, typically leads to the loss of structural cohesion between cells, hence enabling the dissemination of tumour cells in the rest of the organism. Indeen, this gene has been found frequently mutated in invasive breast carcinoma [Lei et al., 2002], in particular in the invasive lobular carcinoma (ILC) where early loss of CDH1 function is not rare. ILC cells usually invade surrounding areas through a specific (non-cohesive) infiltrating behaviour, also called 'Indian-file' [Berx et al., 1995]. Finally, the VHL gene is known to be involved in the degradation of the transcription factor HIF1 which promotes angiogenesis and tumour cell migration. Hence, once mutated, the VHL gene may thus permit the migration of breast cancer cells and the development of surrounding vessels. This provides the migrating tumour cells with an increased glucose intake, which further sustains the primary tumour as well as the formation of metastasis [Zia et al., 2007].

In addition, the network in Fig. 11.1 shows two distinct subnetworks that either involve the genes PTEN/MAP2K4 or the genes PIK3CA/AKT1. As can be seen in Fig. 11.2 [Stephens, 2012], PTEN/MAP2K4 or PIK3CA/AKT1 play a role in

the JUN kinase signalling pathways, located downstream to MAPK8/MAPK9.

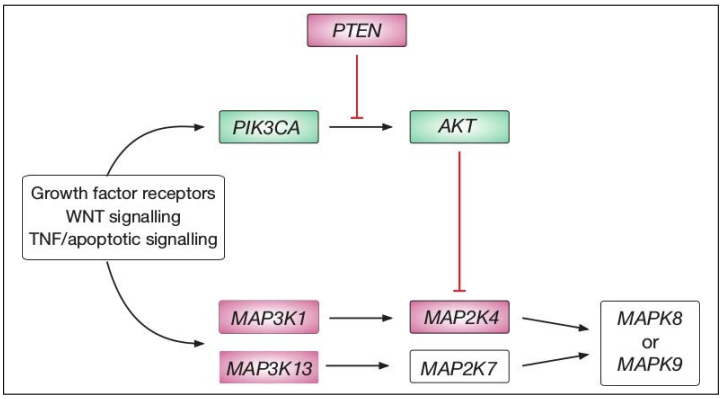


FIGURE 11.2 **JUN kinase signalling pathway.** This schema, taken from [Stephens, 2012], represents the regulation of JUN kinases MAP2K7 and MAP2K8. Genes in green are activated by mutations, while genes in red are inactivated.

The JUN kinases are involved in the cell apoptosis and proliferation processes. The Fig. 11.2 highlights the two following interactions pathways,

$\langle \mathbf{p_A} \rangle$ PIK3CA $\rightarrow$ AKT $-|$ MAPK4 $\rightarrow$ MAPK8 $\rightarrow$ JUN

$\langle \mathbf{p_B} \rangle$ PTEN, which opposes the interaction (PIK3CA $\rightarrow$ AKT) by degrading the proteins encoded by PIK3CA

Following these interactions, previous studies [Stephens, 2012] have predicted that,

- mutations in PIK3CA and/or AKT may induce a constitutive activation that reduce the activity of the JUN pathway
- mutations in PTEN and/or MAP2K4 may induce an inhibition of these genes that also impairs the JUN pathway

Although the reduction of the JUN pathway activity is likely to have complex consequences, it may include the disruption of the cell apoptosis process. The 3off2 reconstruction seems to suggest that impairing either interactions in $\langle \mathbf{p_A} \rangle$ or alternatively in $\langle \mathbf{p_B} \rangle$ is sufficient to trigger and/or sustain the tumour progression by inhibiting the cell apoptosis.

All in all, the 3off2 method could be used to infer temporal cascades of mutations from incomplete observational datasets. Although this study of mutational pathways in breast cancer is still at its early stage, the 3off2 reconstructions already show interesting cascades of alterations. In the future, the inclusion of gene expression levels should also be considered. Indeed, beyond the mutations of the genes themselves, perturbations of signalling pathways through the disruption of regulatory sequences have also a significant impact on the tumour growth that should be incorporated in our analyses.

Chapter 12

Conclusions

This dissertation presents a novel hybrid method, 3off2, for the reconstruction of causal graphical models from finite observational datasets. As shown in the evaluations of Chapter 7, this new approach appears to be more robust than the state-of-the-art methods, namely the constraint-based and Bayesian search-and-score algorithms, when applied to finite observational datasets with inherent sampling noise. As detailed in Chapter 6, the 3off2 hybrid approach benefits from the main advantages of the constraint-based and Bayesian reconstruction methods, and demonstrates a greater robustness than the existing baseline methods, such as the PC-stable algorithm [Colombo and Maathuis, 2014, Spirtes et al., 1993] or the hill-climbing search-and-score heuristics.

Specifically, the 3off2 method combines the Bayesian and constraint-based approaches within a rigorous information-theoretic framework to confidently ascertain structural independencies in causal graphical models underlying *real-life* complex systems. The keystone of the 3off2 approach is the iterative procedure that *takes off* the most likely conditional 3-point information from the 2-point (mutual) information between each pair of nodes to progressively uncover the separation sets and hence learn the direct relationships among the variables (section 6.3.2). The resulting skeleton of this robust iterative process is followed by an orientation/propagation step that relies on the most significant 3-point information terms to eventually provide the direct causal effects of graphical models (sections 6.3.2.2 & 6.4.1).

Earlier hybrid methods have also attempted to improve network reconstruction by combining the concepts of constraint-based approaches with the robustness of Bayesian scores [Cano et al., 2008, Claassen and Heskes, 2012, Dash and Druzdzel, 1999, Tsamardinos et al., 2006]. In particular, [Dash and Druzdzel, 1999] have proposed to exploit an intrinsic weakness of the PC algorithm, its sensitivity to the order in which conditional independencies are tested on finite

data, to rank these different order-dependent PC predictions with Bayesian scores. More recently, [Tsamardinos et al., 2006] have also combined constraint-based and Bayesian approaches by first reconstructing an undirected skeleton through the identification of both parents and children of each nodes of the underlying graphical model and then orienting the edges (and possibly deleting some of them) through a Bayesian hill-climbing heuristics approach restricted to the skeleton. This Max-Min Hill-Climbing (MMHC) approach tends to have a high precision in terms of skeleton but a rather limited sensibility, leading overall to lower skeleton and CPDAG F-scores than 3off2 and Bayesian hill climbing methods on the same benchmark networks (Chapter 7). Interestingly, however, the MMHC approach is among the fastest network reconstruction approaches allowing for scalability to large network sizes [Tsamardinos et al., 2006].

The 3off2 algorithm is expected to run in polynomial time on typical sparse causal networks with low in-degree, just like constraint-based algorithms. However, in practice and despite the additional computation of conditional 2-point and 3-point information terms, we found that the 3off2 algorithm runs typically faster than constraint-based algorithms for large enough samples, due to a cascading accumulation of errors that inflate the combinatorial search of conditional independencies. Instead, we found that 3off2 running time displays a similar trend as Bayesian hill-climbing heuristic methods, (Chapter 7 and Appendices A & B).

As exemplified through Chapters 8 to 11, the 3off2 method can be applied to various biological complex systems, ranging from the transcription factors interplay within a cell (Chapter 10) to the neural activity of the whole brain of a zebrafish larvae (Chapter 9). This could be achieved thanks to the robustness of the 3off2 method that can confidently identify direct causal effects from observational data despite the sampling noise inherent to finite biological datasets. In particular, the 3off2 method can uncover causal graphical models from datasets with missing values, which is a widely encountered issue not only in the field of systems biology but also in epidemiology or social sciences. As introduced in section 6.4.2, the 3off2 method can also allow for the presence of latent variables, an important issue when reconstructing causal graphical models, as one is not typically aware of the whole set of factors or parameters playing a role in the complex process under scrutiny.

The 3off2 method has been extensively evaluated on simulated and *real-life*-like datasets against state-of-the-art methods (Chapter 7) and complementary tests have also been performed in order to evaluate the 3off2 hybrid approach against earlier hybrid methods [Affeldt and Isambert, 2015]. However, the ability of the 3off2 method to assume the influence of hidden variables while learning the causal

network still remains to be extensively evaluated. This will be done by adapting the simulated networks and datasets generated with the causal modeling tool Tetrad IV (section 7.2).

The Chapters 8 to 11 are ongoing projects that are being further investigated in our team. In particular, in order to provide a broader overview of the whole neural activity of the zebrafish larvae brain, we are now designing a 3D clustering method to allow for a 3D causal model reconstruction with the 3off2 approach. This will also require further development in terms of graphical model display within the discoNet pipeline (section 6.5.2). As these 3D reconstructions will clarify the influence of brain areas across several 2D layers at different depths of the brain, this is a nice opportunity to assess the 3off2 method for the assumption of the presence of latent variables, as comparisons will be made possible between the 2D z-plane reconstructions and the 3D global neural network.

Furthermore, our reconstruction of the hematopoietic subnetwork (Chapter 10) relied so far on a dataset that combines single cell-expression data at different stages of the differentiation process. Our team is now exploring the temporal modifications of the hematopoietic subnetwork throughout these distinct stages. The first preliminary results on developmental stage specific datasets already highlight well-known hubs among the transcription factors playing a role in the differentiation of the stem cells. Besides, the 3off2 approach also indicates sensible inhibitions and activations that are turned on and off along this complex biological process.

Importantly, the 3off2 method can be used for multivariate analysis that integrate heterogeneous datasets. Hence, as introduced in Chapter 11, complementary factors will be included in the analysis of tumour progression to provide a broader understanding of the tumour emergence and spreading process. Indeed, beyond single point mutations, we are now considering the role of the copy number variations and the expression levels of major drivers identified in several cancer types. Whenever available, we will also considered complementary factors, such as clinical information (*e.g.* survival time, patient age, tumour stage).

Finally, a long-term development goal is to guarantee that this novel 3off2 approach for complex (biological) network reconstructions can be scaled up to thousands of variables. These future developments, that require several adaptations of the 3off2 robust scheme, are now under study and design.

Appendix A

Complementary evaluations on *real-life* networks

In this appendix, the results of the considered network reconstruction methods are evaluated using different parameters values and against each others in terms of Precision (or positive predictive value), $Prec = TP/(TP + FP)$, Recall or Sensitivity (true positive rate), $Rec = TP/(TP + FN)$, as well as F-score $= 2 \times Prec \times Rec/(Prec + Rec)$ and execution time when comparing the CPDAG (filled lines) of the reconstructed network (or its skeleton (dashed line)) to the CPDAG (or the skeleton) of the benchmark network. The alternative methods are:

- the stable version of the PC algorithm [Colombo and Maathuis, 2014, Spirtes and Glymour, 1991] implemented in the `pcalg` package [Kalisch and Bühlmann, 2008, Kalisch et al., 2012]
- the Bayesian inference method using the hill-climbing heuristics implemented in the `bnlearn` package [Scutari, 2010]
- Aracne [Margolin et al., 2006], an information-based inference approach, which iteratively prunes links with the weakest mutual information based on the Data Processing Inequality. We have used the Aracne implementation of the `minet` package [Meyer et al., 2008]

For sample sizes from $N = 10$ to $50,000$ data points, the methods have been tested on 50 replicates and the Figures A.1-A.26 give the average results over these multiple replicates. The five following benchmark networks have been considered (refer to the `bnlearn` package [Scutari, 2010] for more details on these networks):

CHILD 20 nodes, 25 links, 230 parameters, mean degree 2.5, maximum in-degree 2

INSURANCE 27 nodes, 52 links, 984 parameters, mean degree 3.85, maximum in-degree 3

ALARM 37 nodes, 46 links, 509 parameters, mean degree 2.49, maximum in-degree 4

BARLEY 48 nodes, 84 links, 114,005 parameters, mean degree 3.5, maximum in-degree 4

HEPAR II 70 nodes, 123 links, 1,453 parameters, mean degree 3.51, maximum in-degree 6

A.1 Evaluation of the PC method by significance level

In this section, the results of the PC inference method [Spirtes and Glymour, 1991], as implemented in the `pcalg` package [Kalisch and Bühlmann, 2008, Kalisch et al., 2012], are evaluated using the following parameters values:

- a significance level $\alpha = 0.001$ (**PC 1e-03**)
- a significance level $\alpha = 0.01$ (**PC 1e-02**)
- a significance level $\alpha = 0.1$ (**PC 1e-01**)

In particular, the *stable* implementation of the PC algorithm has been used, as well as the *majority rule* for the orientation and propagation steps [Colombo and Maathuis, 2014]. As shown in Figures A.1-A.5, the PC inference method typically requires the significance level to be adjusted to larger values ($\alpha = 0.1$) at small sample sizes and to smaller values ($\alpha = 0.01$ or $\alpha = 0.001$) on larger datasets to improve the CPDAG F-scores.

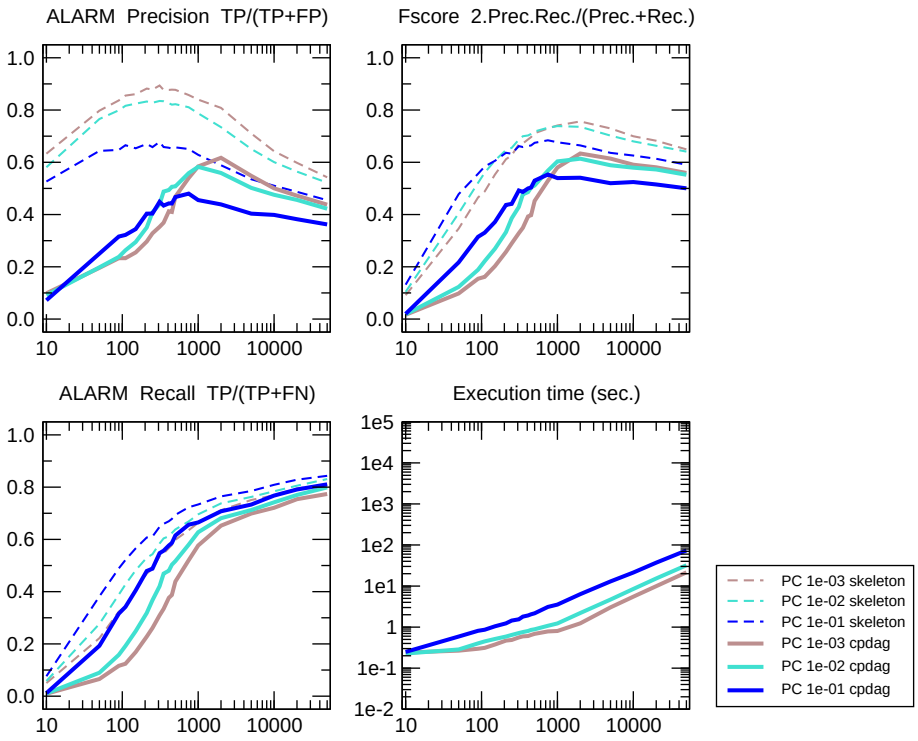


FIGURE A.1 **ALARM network** [37 nodes, 46 links, 509 parameters, Average degree 2.49, Maximum in-degree 4]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using the PC inference approach.

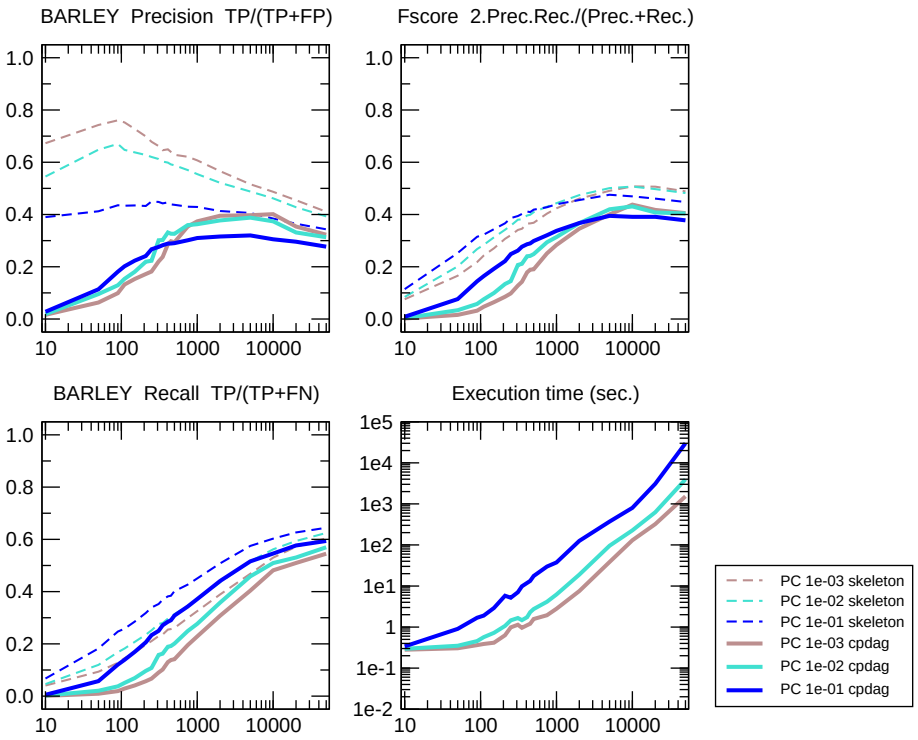


FIGURE A.2 **BARLEY network** [48 nodes, 84 links, 114,005 parameters, Average degree 3.5, Maximum in-degree 4]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using the PC inference approach.

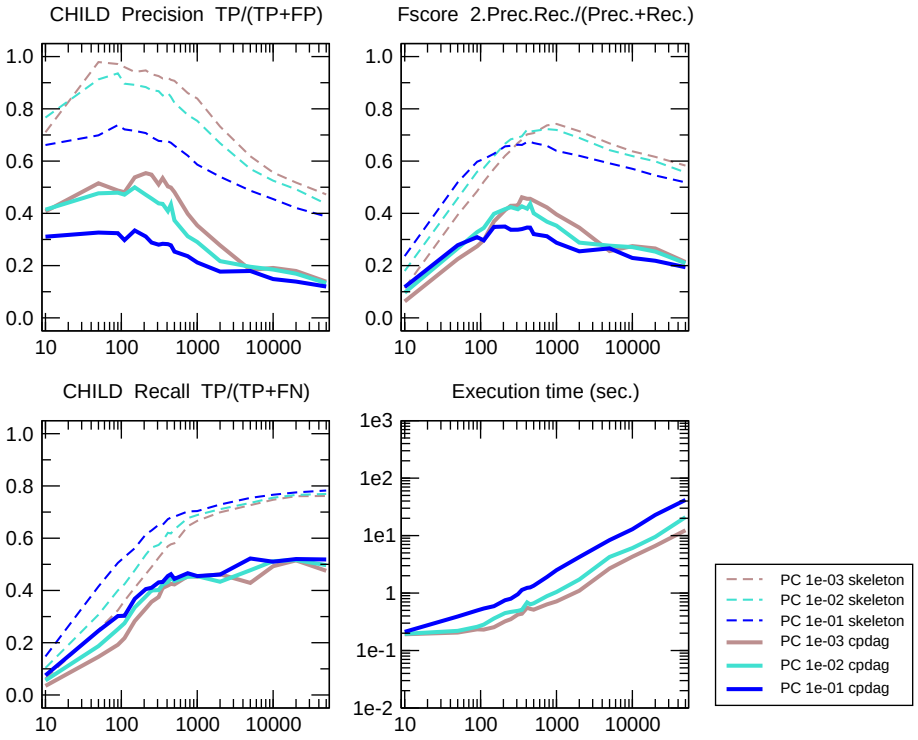


FIGURE A.3 **CHILD network** [20 nodes, 25 links, 230 parameters, Average degree 2.5, Maximum in-degree 2]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using the PC inference approach.

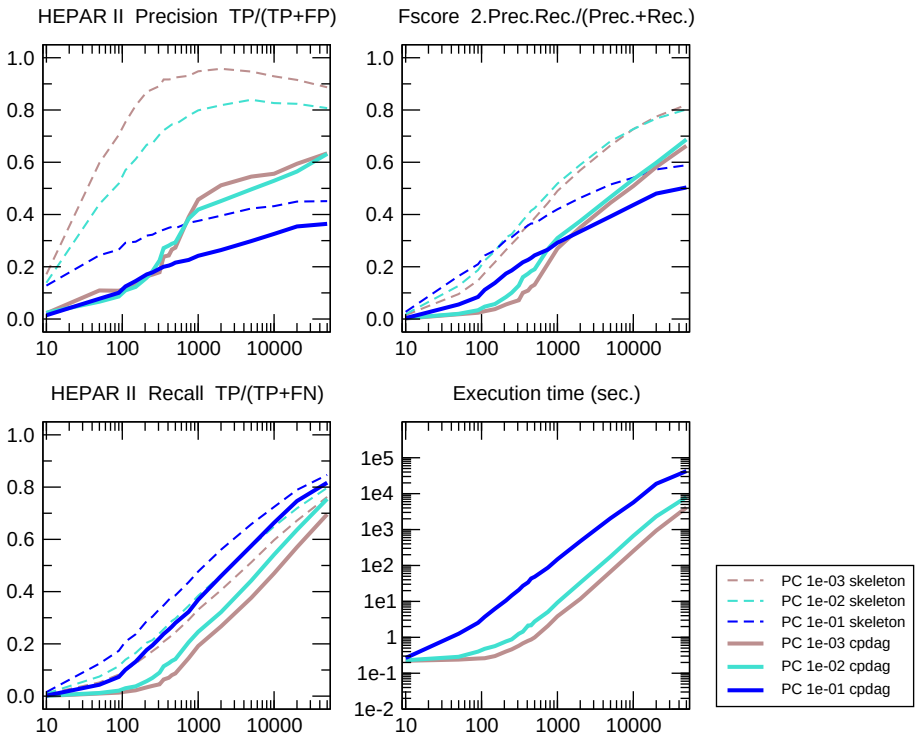


FIGURE A.4 **HEPAR II network** [70 nodes, 123 links, 1,453 parameters, Average degree 3.51, Maximum in-degree 6]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using the PC inference approach.

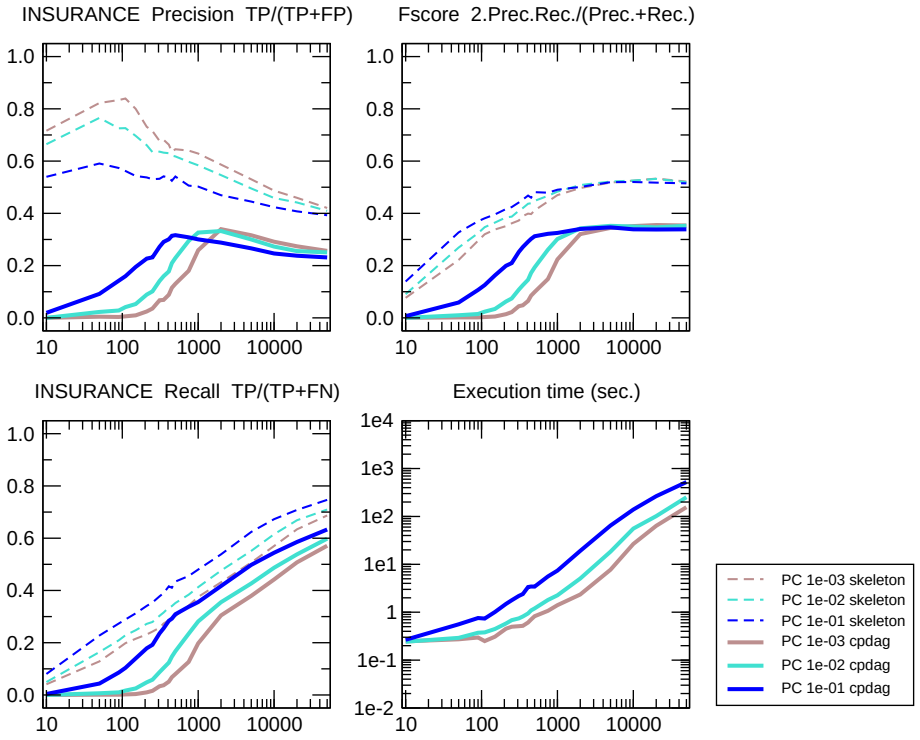


FIGURE A.5 **INSURANCE network** [27 nodes, 52 links, 984 parameters, Average degree 3.85, Maximum in-degree 3]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using the PC inference approach.

A.2 Evaluation of the Aracne reconstruction method

In this section, the results of the Aracne inference method, as implemented in the `minet` package [Meyer et al., 2008], are evaluated using the following parameters values:

- a threshold for the minimum difference in mutual information set to $\epsilon = 1/N$ (**Aracne Eps 1/N**)
- a threshold for the minimum difference in mutual information set to $\epsilon = 0$ (**Aracne Eps 0**)

As shown in Figures A.6-A.10, small positive values for the threshold parameters for minimum difference in mutual information typically worsen F-scores.

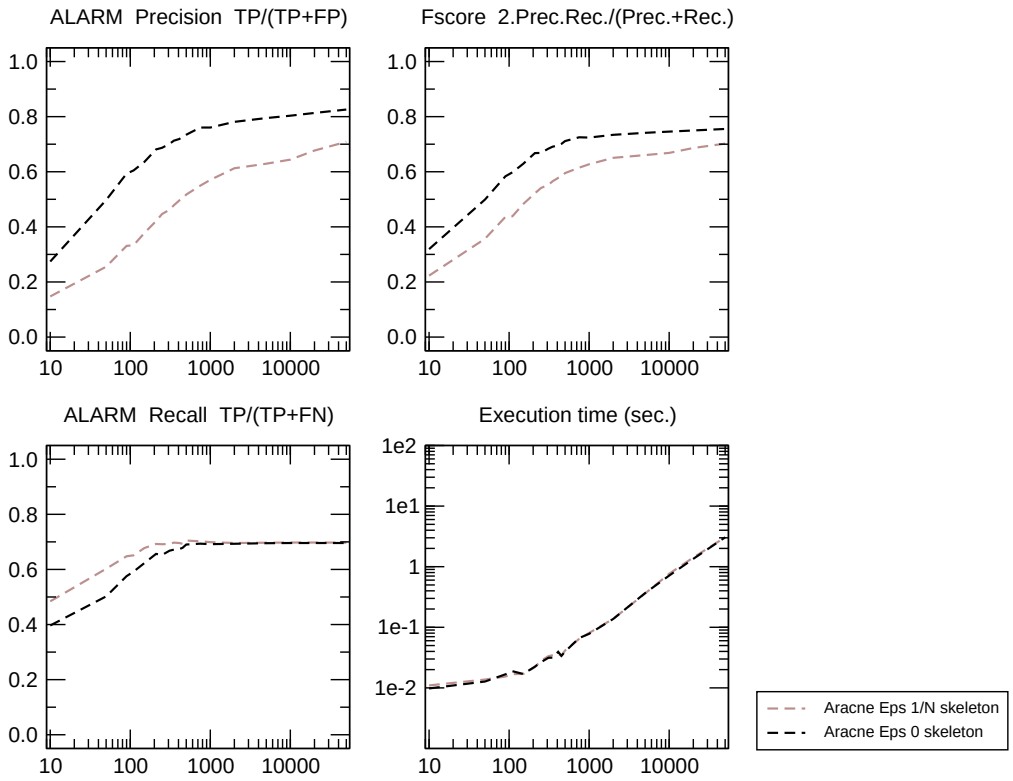


FIGURE A.6 **ALARM network** [37 nodes, 46 links, 509 parameters, Average degree 2.49, Maximum in-degree 4]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) using the Aracne inference approach.

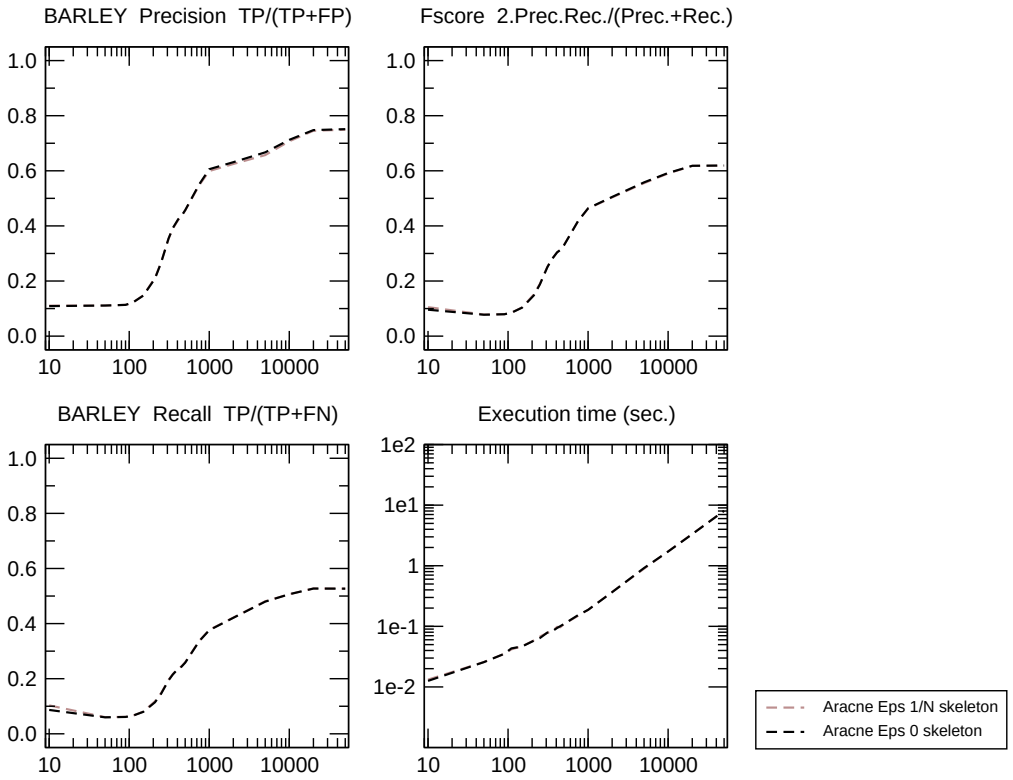


FIGURE A.7 **BARLEY network** [48 nodes, 84 links, 114,005 parameters, Average degree 3.5, Maximum in-degree 4]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) using the Aracne inference approach.

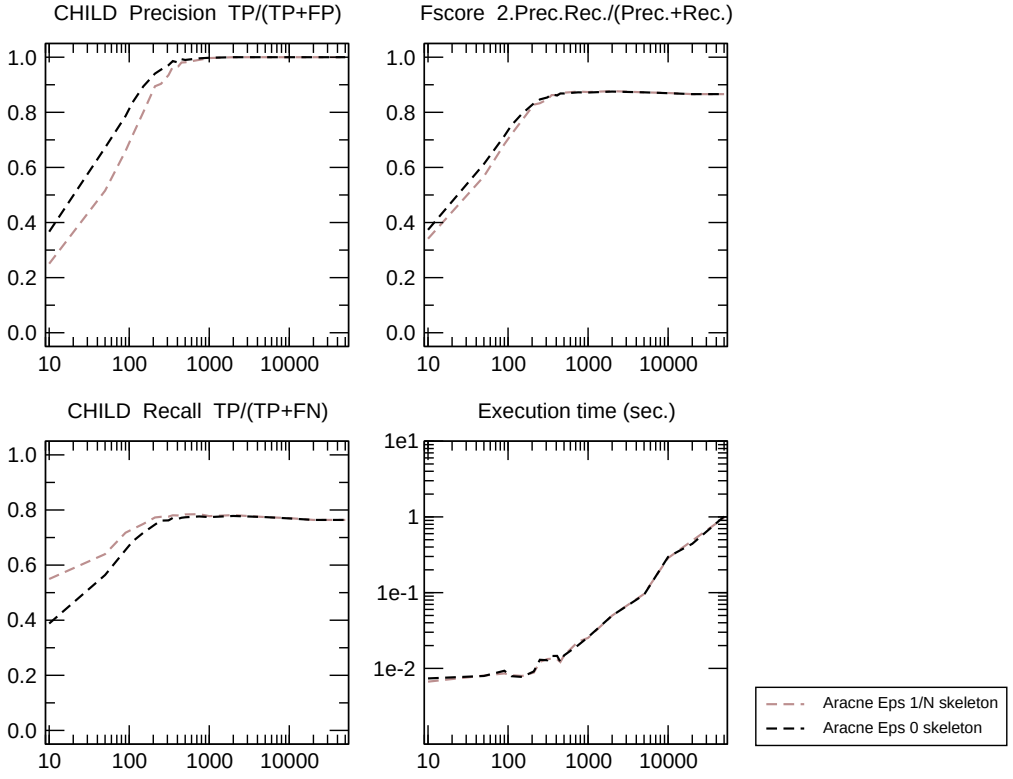


FIGURE A.8 **CHILD network** [20 nodes, 25 links, 230 parameters, Average degree 2.5, Maximum in-degree 2]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) using the Aracne inference approach.

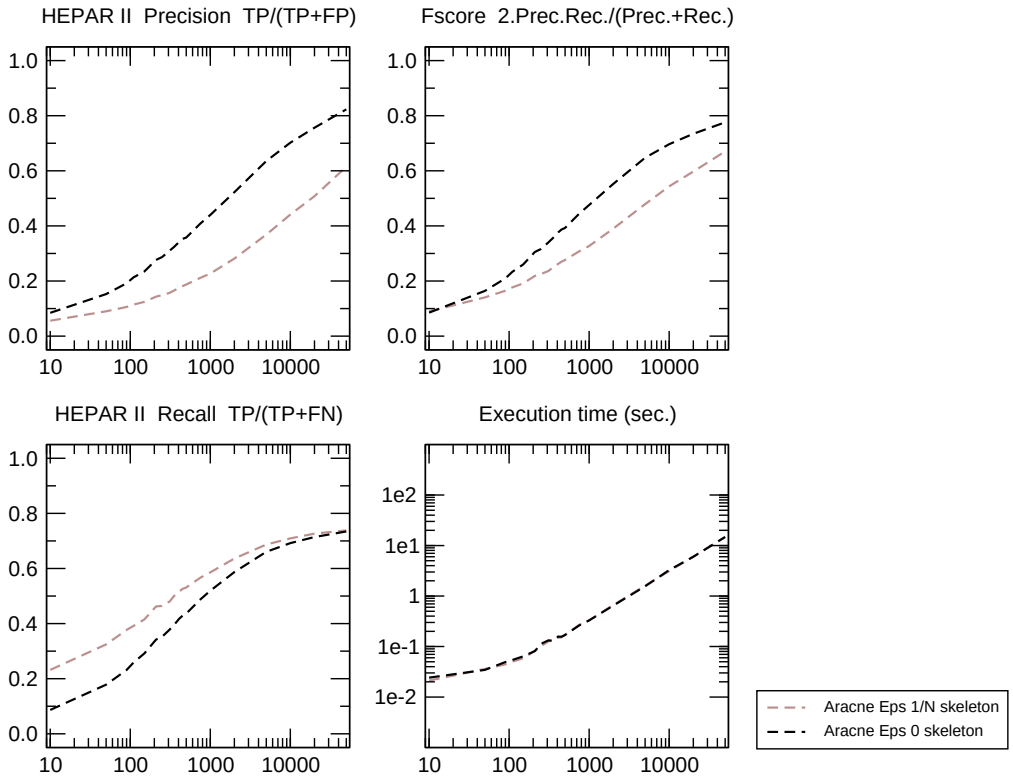


FIGURE A.9 **HEPAR II network** [70 nodes, 123 links, 1,453 parameters, Average degree 3.51, Maximum in-degree 6]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) using the Aracne inference approach.

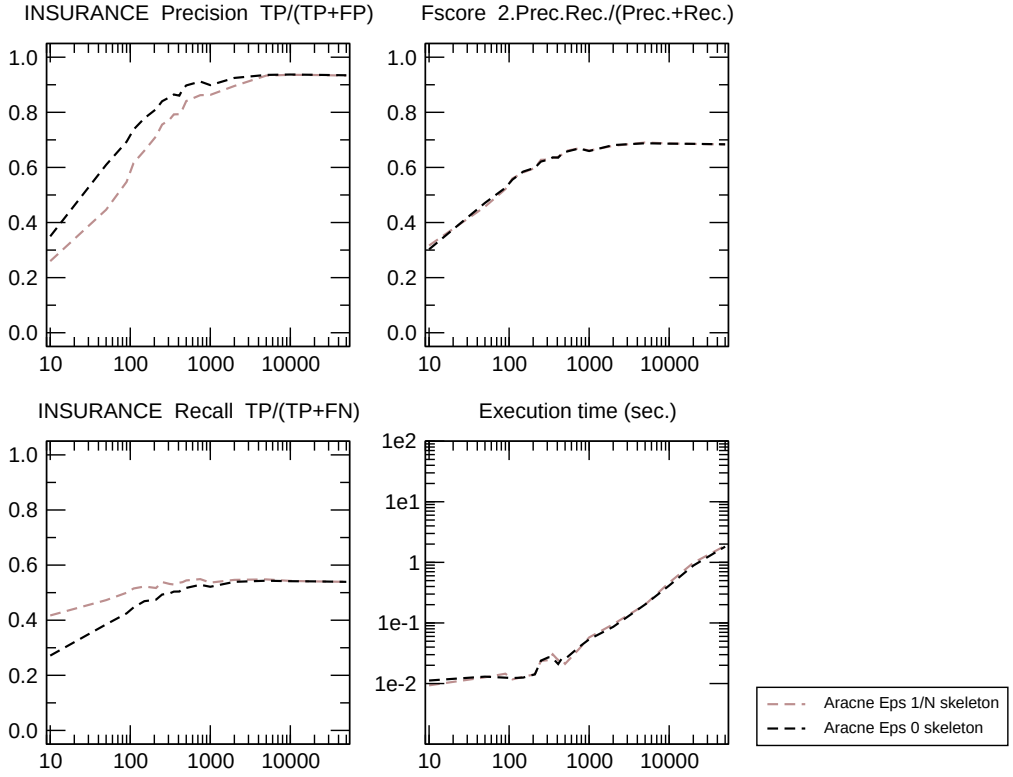


FIGURE A.10 **INSURANCE network** [27 nodes, 52 links, 984 parameters, Average degree 3.85, Maximum in-degree 3]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) using the Aracne inference approach.

A.3 Evaluation of the Bayesian methods by score

In this section, the results of the Bayesian inference method using a hill-climbing (HC) heuristics with 100 random restarts [Chickering et al., 1995], as implemented in the `bnlearn` package [Scutari, 2010], are evaluated using the following parameters values:

- Bayesian Dirichlet equivalent (BDe) score (**HC BDe**)
- Akaike Information Criterion (AIC) (**HC AIC**)
- Bayesian Information Criterion (BIC) score (**HC BIC**)

As shown in Figures A.11-A.15, depending on the underlying causal network, one has to choose the most suitable score for the hill-climbing heuristic approach to output its best reconstruction. The AIC score should be preferred for INSURANCE (Figure A.15), the BIC score for the ALARM and HEPAR II (Figures A.11 & A.14) and the BDe score for the CHILD and BARLEY networks (Figures A.13 & A.12).

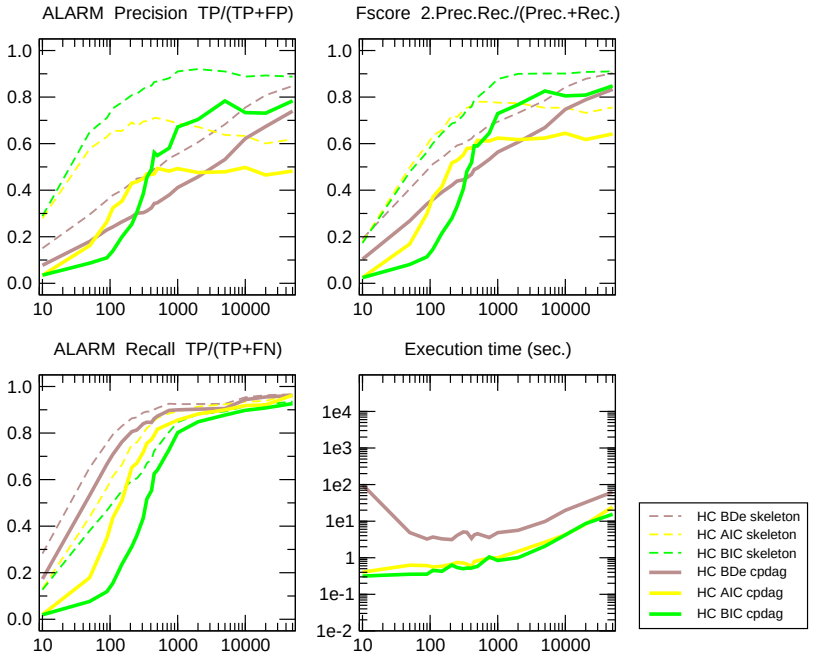


FIGURE A.11 **ALARM network** [37 nodes, 46 links, 509 parameters, Average degree 2.49, Maximum in-degree 4]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using the Bayesian inference approach with BDe, AIC or BIC scores. The Bayesian inference method using the hill climbing heuristics and the BDe score didn't converge for 4 datasets out of 50 at sample size $N = 10$. For these non-converging reconstructions, the execution time has been set to $3,600$ seconds.

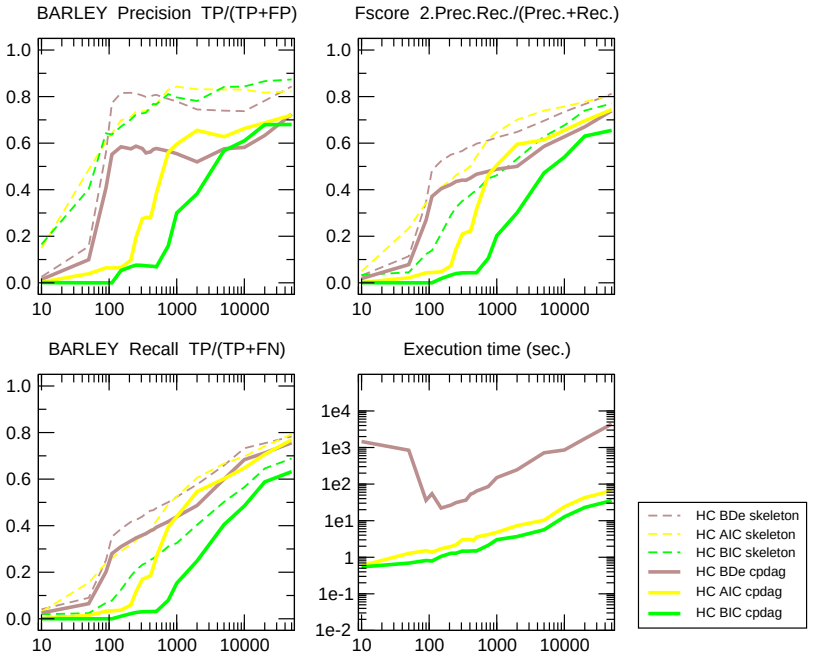


FIGURE A.12 **BARLEY network** [48 nodes, 84 links, 114,005 parameters, Average degree 3.5, Maximum in-degree 4]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using the Bayesian inference approach with BDe, AIC or BIC scores. The Bayesian inference method using the hill climbing heuristics and the BDe score didn't converge for (i) 41 datasets out of 50 at sample size $N = 10$, (ii) 36 datasets out of 50 at sample size $N = 50$, (iii) 13 datasets out of 50 at sample size $N = 90$ and (iv) 1 dataset out of 50 at sample size $N = 110$. For these non-converging reconstructions, the execution time has been set to $3,600$ seconds.

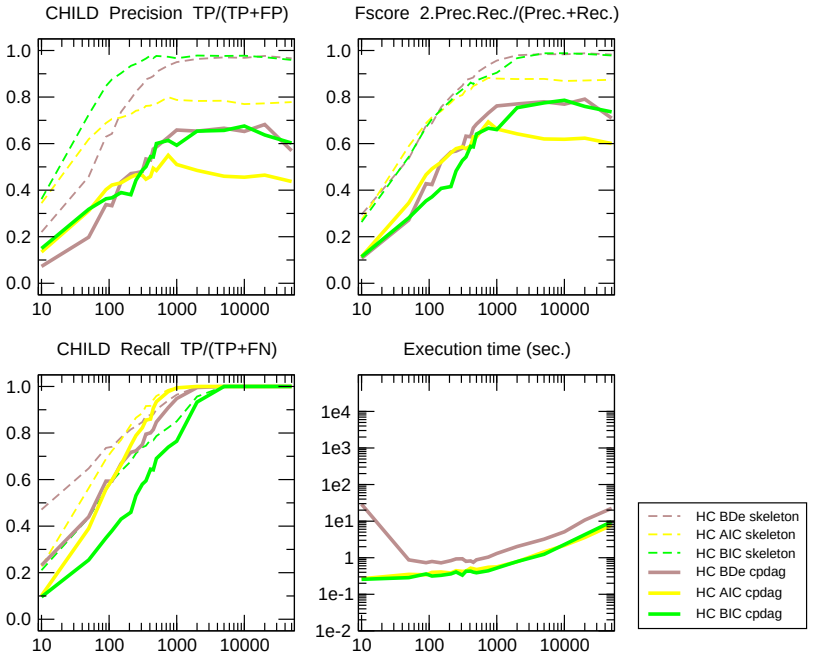


FIGURE A.13 **CHILD network** [20 nodes, 25 links, 230 parameters, Average degree 2.5, Maximum in-degree 2]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using the Bayesian inference approach with BDe, AIC or BIC scores.

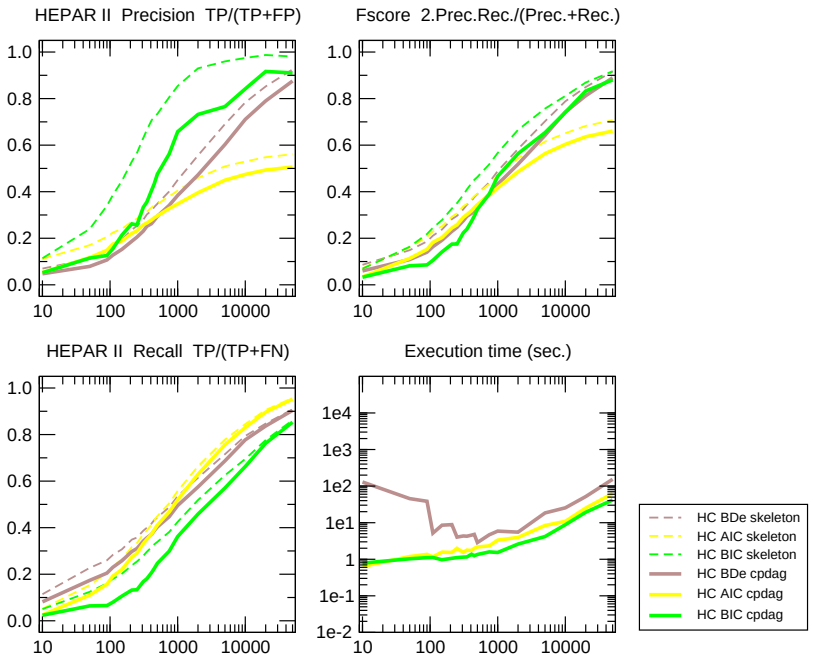


FIGURE A.14 **HEPAR II network** [70 nodes, 123 links, 1,453 parameters, Average degree 3.51, Maximum in-degree 6]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using the Bayesian inference approach with BDe, AIC or BIC scores. The Bayesian inference method using the hill climbing heuristics and the BDe score didn't converge for 5 datasets out of 50 at sample size $N = 10$. For these non-converging reconstructions, the execution time has been set to $3,600$ seconds.

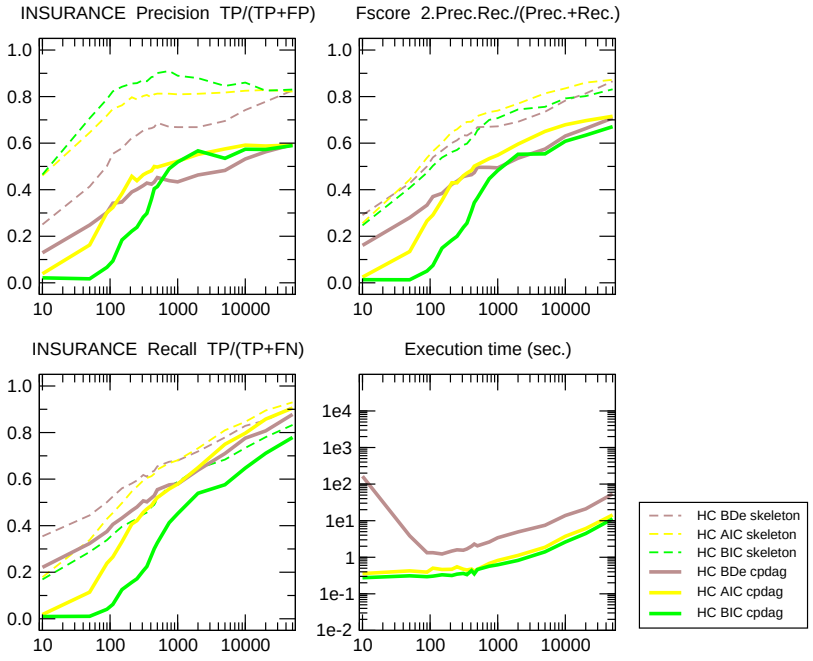


FIGURE A.15 **INSURANCE network** [27 nodes, 52 links, 984 parameters, Average degree 3.85, Maximum in-degree 3]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using the Bayesian inference approach with BDe, AIC or BIC scores.

A.4 Evaluation of 3off2 by score

In this section, the results of the 3off2 inference approach are evaluated using the following parameters values:

- 3off2 **MDL(rank I)**: 3off2 inference approach using Minimum Description Length (MDL) criterion and *non-shifted* 2-point and 3-point information terms in the rank of individual edges
- 3off2 **MDL(rank I')**: 3off2 inference approach using Minimum Description Length (MDL) criterion and *shifted* 2-point and 3-point information terms in the rank of individual edges
- 3off2 **NML(rank I)**: 3off2 inference approach using Normalized Maximum Likelihood (NML) criterion and *non-shifted* 2-point and 3-point information terms in the rank of individual edges
- 3off2 **NML(rank I')**: 3off2 inference approach using Normalized Maximum Likelihood (NML) criterion and *shifted* 2-point and 3-point information terms in the rank of individual edges

As discussed in section 6.3.2, the Figures A.16-A.20 show that the MDL complexity using shifted 2-point and 3-point information terms leads the 3off2 inference approach to cumulate false negative edges with smaller datasets. Instead, the 3off2 inference approach using the NML criterion with shifted 2-point and 3-point information demonstrates very good results for all sample sizes, and for the five benchmark networks.

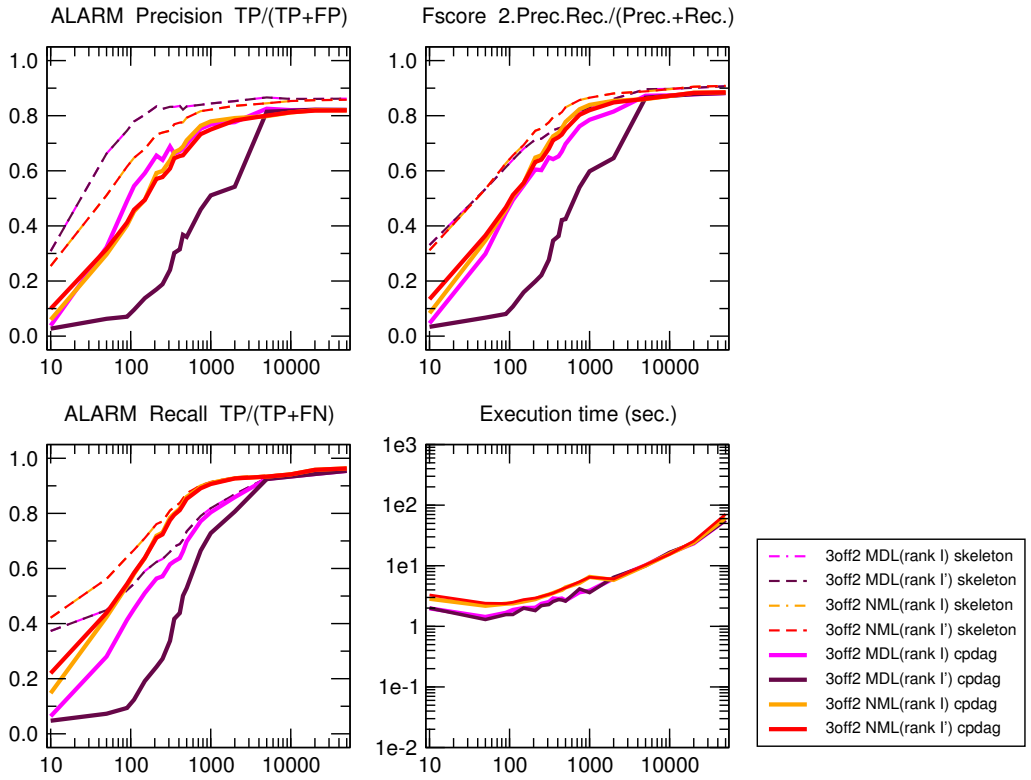


FIGURE A.16 **ALARM network** [37 nodes, 46 links, 509 parameters, Average degree 2.49, Maximum in-degree 4]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using the 3off2 inference approach.

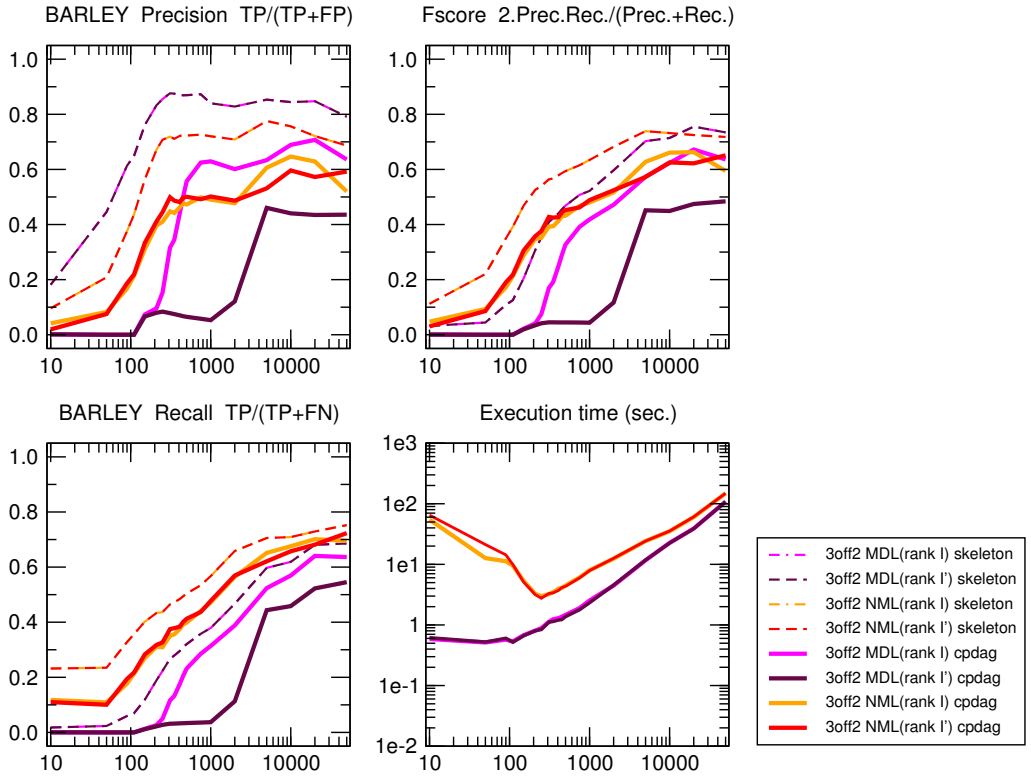


FIGURE A.17 **BARLEY network** [48 nodes, 84 links, 114,005 parameters, Average degree 3.5, Maximum in-degree 4]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using the 3off2 inference approach.

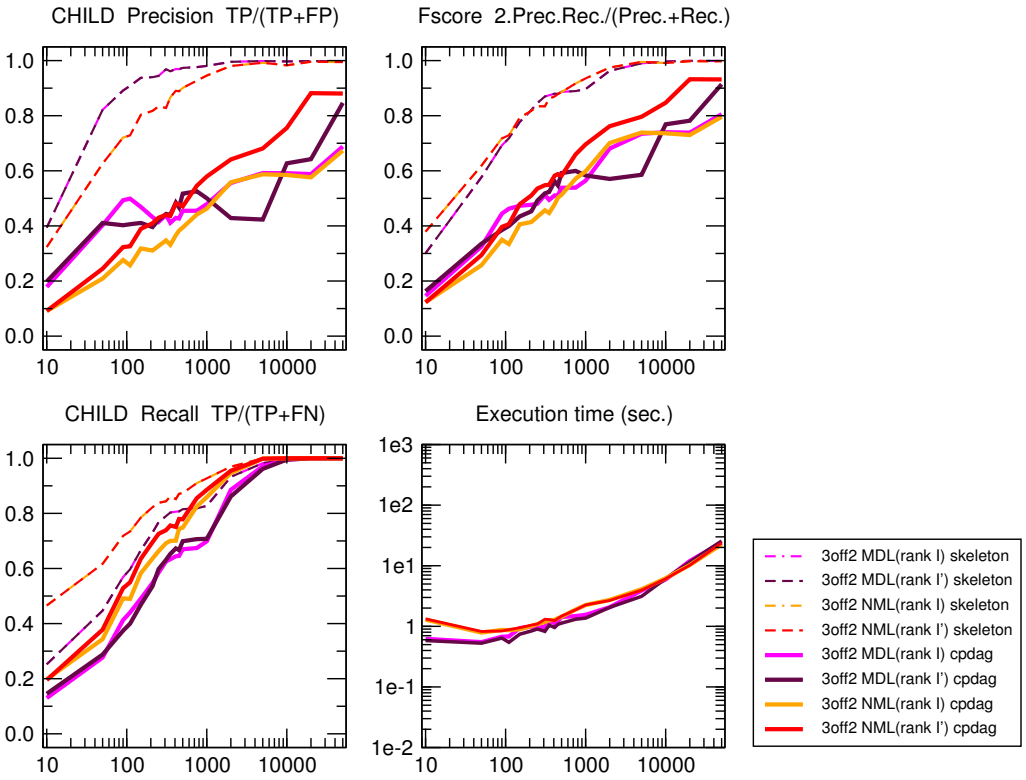


FIGURE A.18 **CHILD network** [20 nodes, 25 links, 230 parameters, Average degree 2.5, Maximum in-degree 2]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using the 3off2 inference approach.

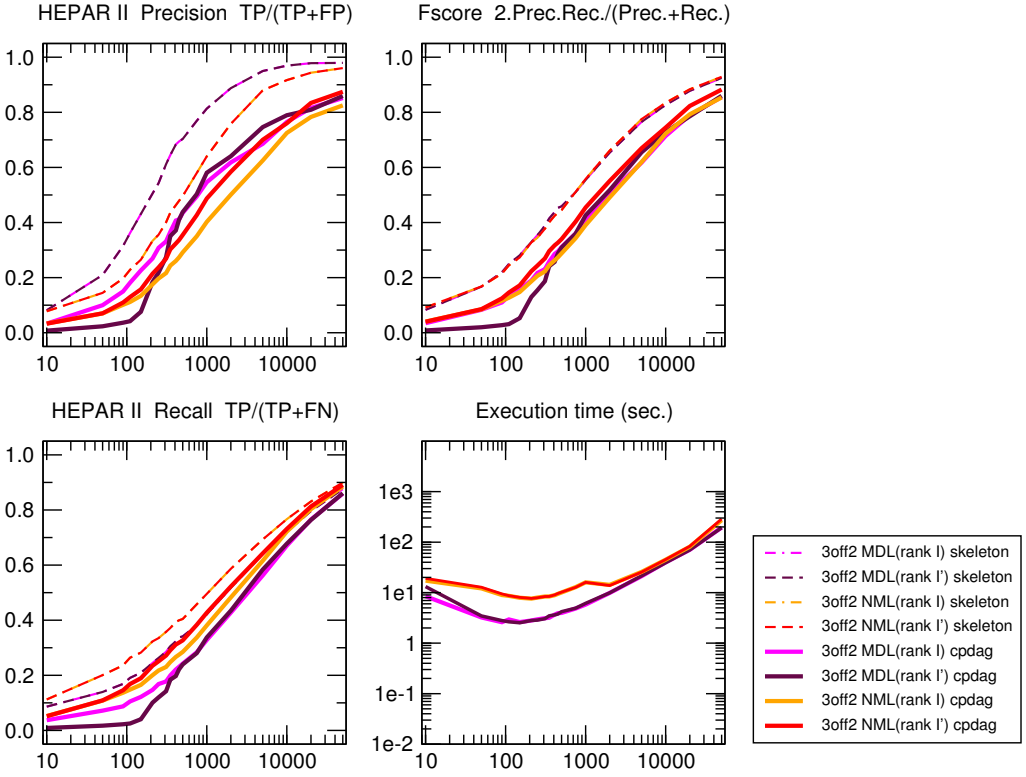


FIGURE A.19 **HEPAR II network** [70 nodes, 123 links, 1,453 parameters, Average degree 3.51, Maximum in-degree 6]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using the 3off2 inference approach.

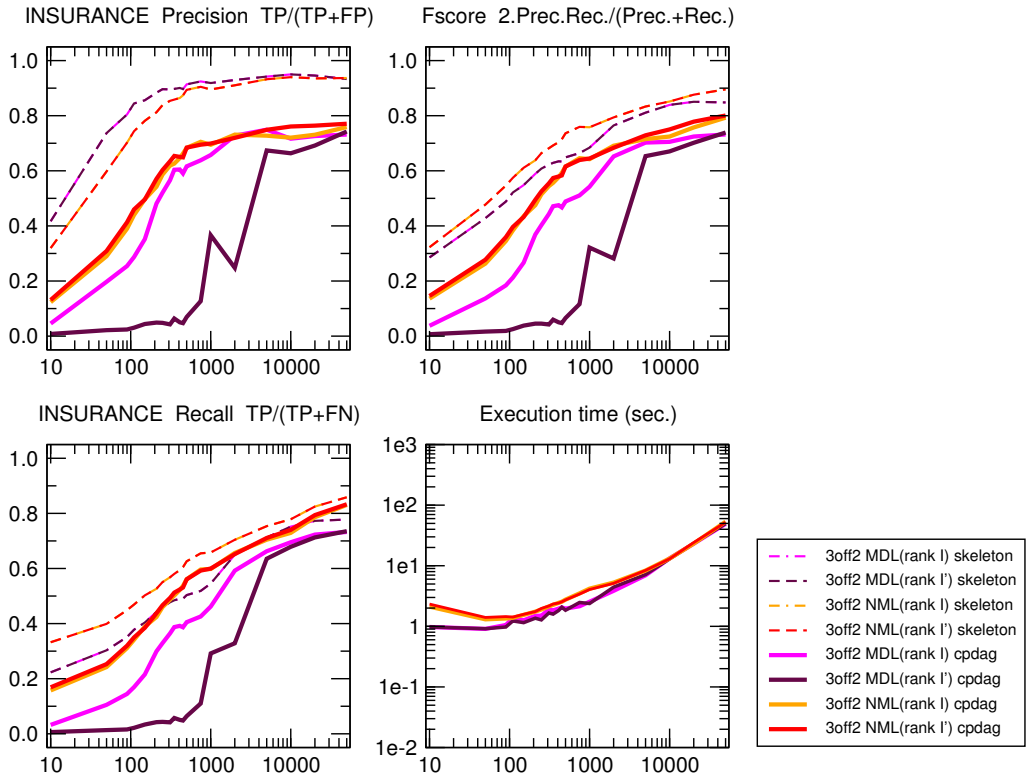


FIGURE A.20 **INSURANCE network** [27 nodes, 52 links, 984 parameters, Average degree 3.85, Maximum in-degree 3]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using the 3off2 inference approach.

A.5 Evaluation of 3off2 against Bayesian and MMHC methods

In this section, the results of the 3off2 inference approach are evaluated against the Bayesian inference methods and a hybrid method using the following parameters values:

- **HC BIC**: Bayesian inference using Bayesian Information Criterion (BIC) score and hill-climbing (HC) heuristics with 100 random restarts [Chickering et al., 1995] as implemented in the `bnlearn` package [Scutari, 2010]
- MMHC **BIC**($\alpha = 1e^{-1}$): Hybrid approach [Tsamardinos et al., 2006] reconstructing first an undirected skeleton through the identification of the parents and children of each node of the underlying graph and then orienting the edges of the skeleton through a Bayesian hill climbing heuristics (possibly removing edges) using the BIC criterion with significance parameter $\alpha = 0.1$ (using BDe criterion and the range $\alpha = 0.001 - 0.1$ gives very similar results for all tested benchmarks, not shown). The MMHC method is implemented in the `bnlearn` package [Scutari, 2010].

- 3off2 **MDL(rank I)**: 3off2 inference approach using Minimum Description Length (MDL) criterion and *non-shifted* 2-point and 3-point information terms in the rank of individual edges

As discussed in section 6.3.2, the MDL complexity using shifted 2-point and 3-point information terms leads the 3off2 inference approach to cumulate false negative edges with smaller datasets. Yet, the 3off2 algorithm gives good results when using the MDL criterion with *non-shifted* 2-point and 3-point information terms in the rank of individual edges. As shown in Figures A.21-A.25, CPDAG F-scores of the 3off2 reconstruction method are typically better or comparable to the Bayesian hill-climbing heuristics using the BIC score, except for the CHILD benchmark network, although 3off2 eventually reaches better results than the Bayesian inference appraoch for very large datasets ($N > 20,000$).

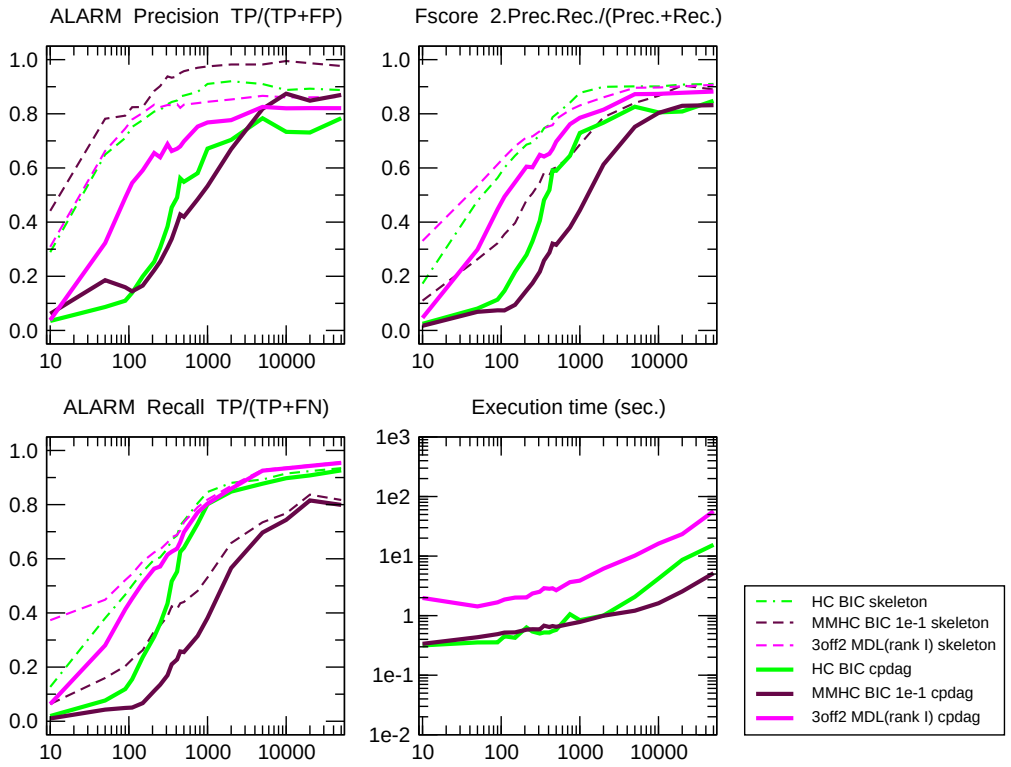


FIGURE A.21 **ALARM network** [37 nodes, 46 links, 509 parameters, Average degree 2.49, Maximum in-degree 4]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using 3off2, Bayesian hill-climbing and Max-Min Hill-Climbing approaches.

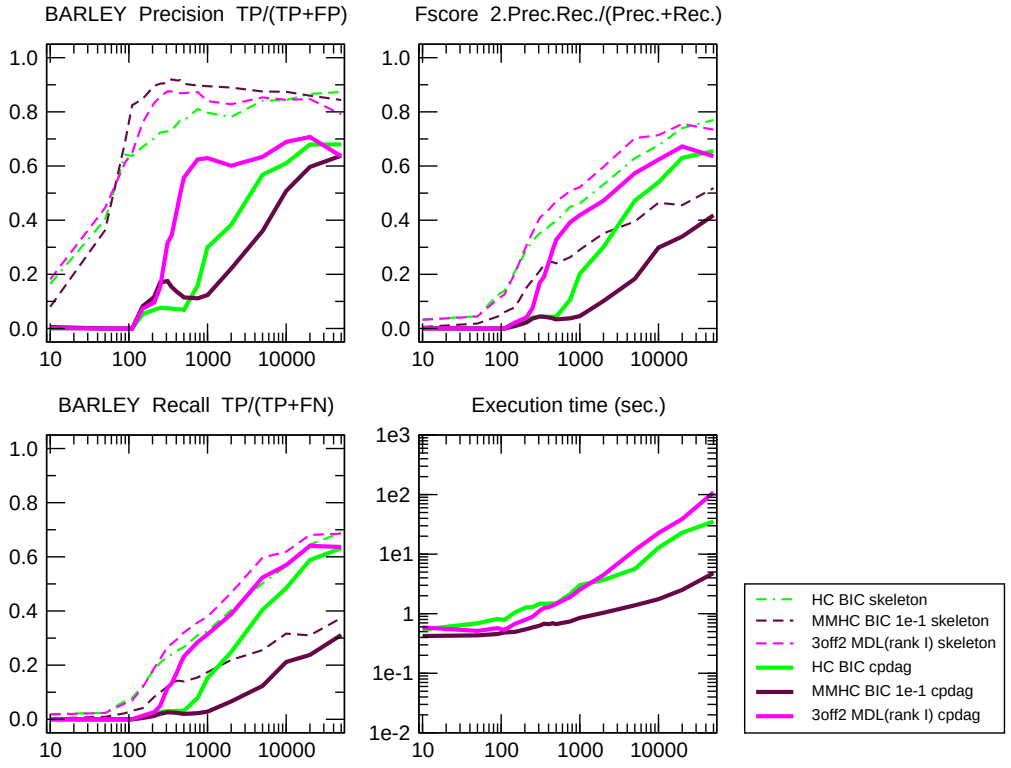


FIGURE A.22 **BARLEY network** [48 nodes, 84 links, 114,005 parameters, Average degree 3.5, Maximum in-degree 4]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using 3off2, Bayesian hill-climbing and Max-Min Hill-Climbing approaches.

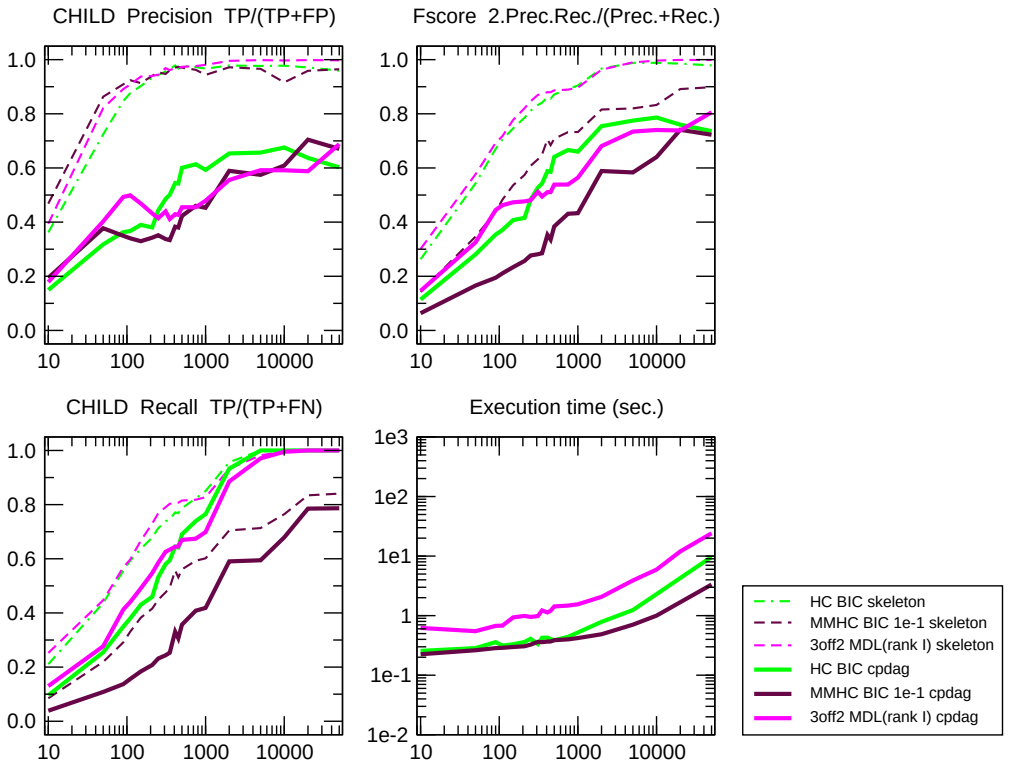


FIGURE A.23 **CHILD network** [20 nodes, 25 links, 230 parameters, Average degree 2.5, Maximum in-degree 2]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using 3off2, Bayesian hill-climbing and Max-Min Hill-Climbing approaches.

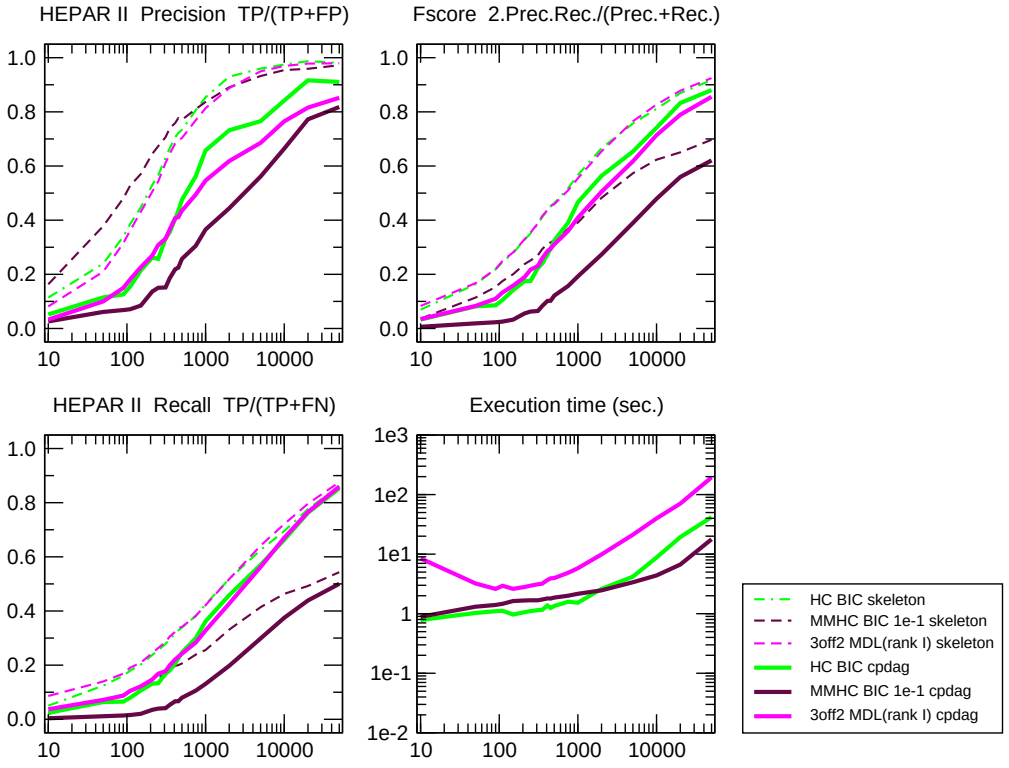


FIGURE A.24 **HEPAR II network** [70 nodes, 123 links, 1,453 parameters, Average degree 3.51, Maximum in-degree 6]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using 3off2, Bayesian hill-climbing and Max-Min Hill-Climbing approaches.

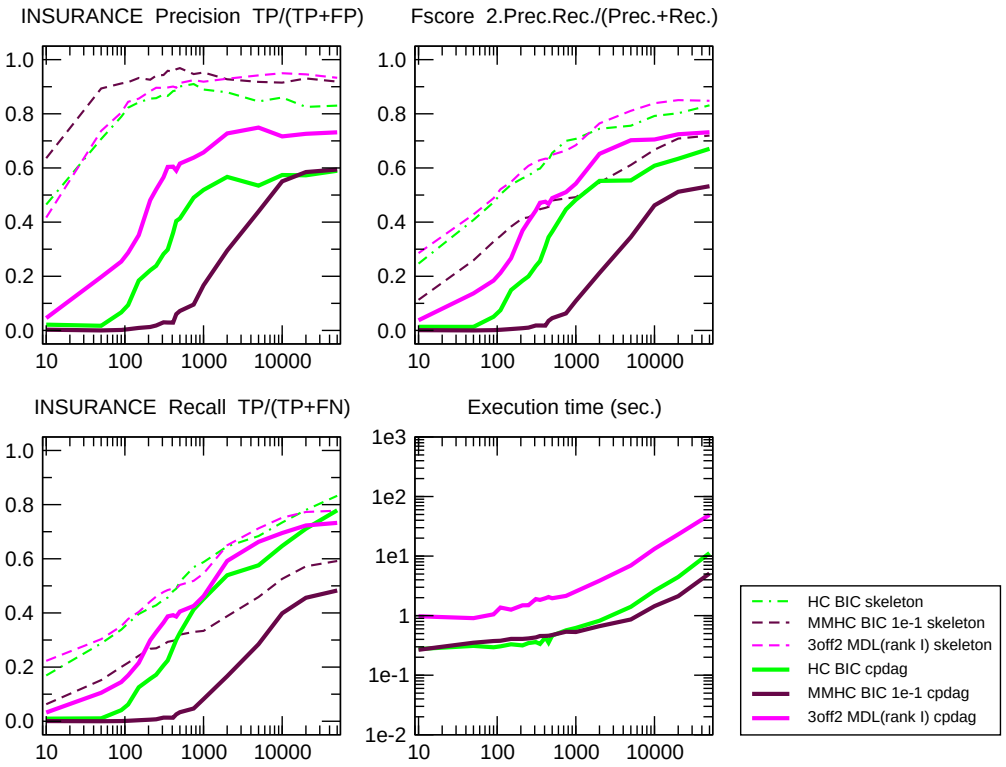


FIGURE A.25 **INSURANCE network** [27 nodes, 52 links, 984 parameters, Average degree 3.85, Maximum in-degree 3]. Precision, Recall, F-score and execution time for the reconstruction of the skeletons (dashed lines) and CPDAGs (filled lines) using 3off2, Bayesian hill-climbing and Max-Min Hill-Climbing approaches.

A.6 Execution time comparisons

This section is comparing the execution time of the different inference methods when reconstructing the CPDAG of the causal benchmark network that has generated the datasets. The methods and parameters values are:

- 3off2 **MDL(rank I)**: 3off2 inference approach using Minimum Description Length (MDL) criterion and *non-shifted* 2-point and 3-point information terms in the rank of individual edges
- 3off2 **NML(rank I')**: 3off2 inference approach using Normalized Maximum Likelihood (NML) criterion and *shifted* 2-point and 3-point information terms in the rank of individual edges
- **PC 1e-01**: PC inference method [Spirtes and Glymour, 1991] implemented in the `pcalg` package [Kalisch and Bühlmann, 2008, Kalisch et al., 2012] with a significance level $\alpha = 0.1$
- **HC BIC**: Bayesian inference using Bayesian Information Criterion (BIC) score and hill-climbing (HC) heuristics with 100 random restarts [Chickering et al., 1995] implemented in the `bnlearn` package [Scutari, 2010]

- MMHC **BIC**($\alpha = 1e^{-1}$): Hybrid approach [Tsamardinos et al., 2006] reconstructing first an undirected skeleton through the identification of the parents and children of each node of the underlying graph and then orienting the edges of the skeleton through a Bayesian hill climbing heuristics (possibly removing edges) using the BIC criterion with significance parameter $\alpha = 0.1$ (using BDe criterion and the range $\alpha = 0.001 - 0.1$ gives very similar results for all tested benchmarks, not shown). The MMHC method is implemented in the `bnlearn` package [Scutari, 2010].

As shown in Figure A.26, the 3off2 reconstruction method is typically fast for the five benchmark networks. In particular, although the PC algorithm is faster at small sample sizes, the 3off2 inference approach becomes rapidly faster on larger datasets. This could be explain by the accumulation of errors during the reconstruction of the PC inference approach, leading to perform too many statistical tests on false positive edges.

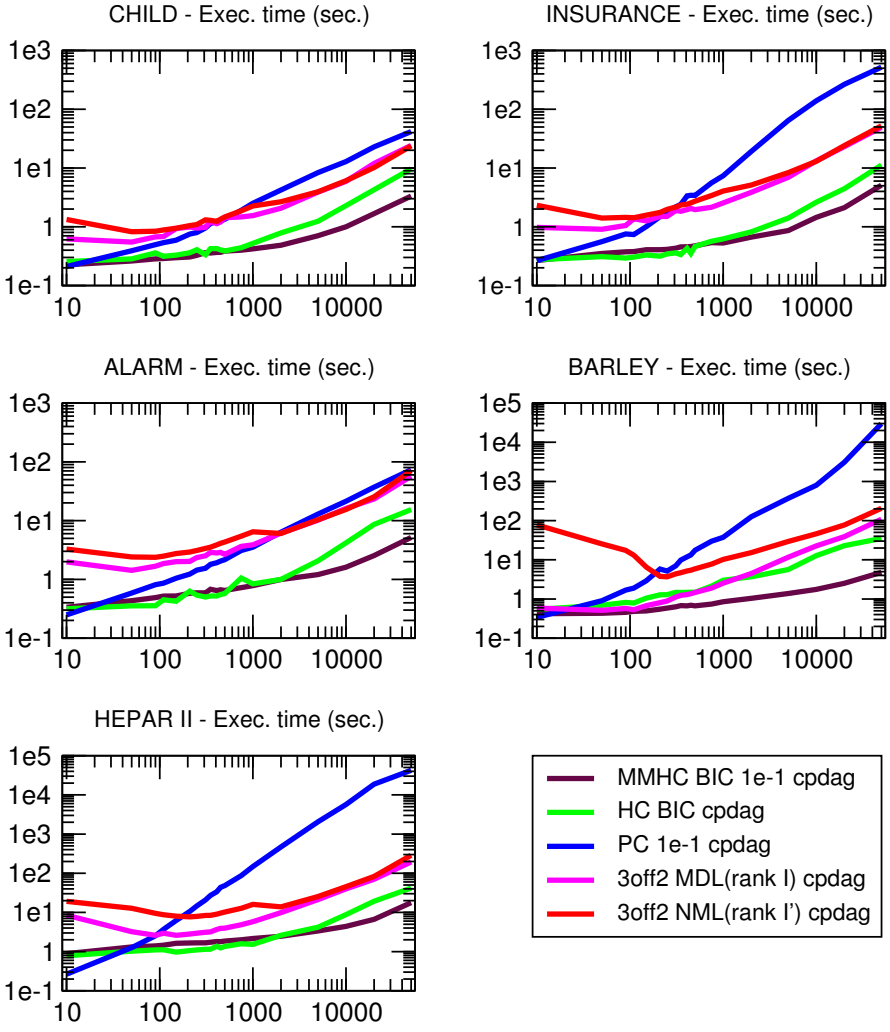


FIGURE A.26 **Network Reconstruction Execution time** Execution time in seconds for the reconstruction of the five studied benchmark networks.

Appendix B

Complementary evaluations on simulated networks

B.1 Description of the benchmark networks

nodes	edges	$\langle k \rangle$	Model	$\langle k_{\max} \rangle$	$\langle k^{in}_{\max} \rangle$	$\langle k^{out}_{\max} \rangle$	N	Replicates
50	**20**	0.8	*1*	4	2	3	$[50 - 50,000]$	20
			2	4	2	2	$[50 - 50,000]$	20
			3	3	3	2	$[50 - 50,000]$	20
			4	3	3	2	$[50 - 50,000]$	20
			5	3	2	2	$[50 - 50,000]$	20
			Avg.	**3.4**	**2.4**	**2.2**		
50	**40**	1.6	*1*	5	3	5	$[50 - 50,000]$	20
			2	6	3	3	$[50 - 50,000]$	20
			3	5	3	3	$[50 - 50,000]$	20
			4	4	4	4	$[50 - 50,000]$	20
			5	5	3	3	$[50 - 50,000]$	20
			Avg.	**5**	**3.2**	**3.6**		
50	**60**	2.4	*1*	7	5	3	$[50 - 50,000]$	20
			2	6	6	3	$[50 - 50,000]$	20
			3	6	4	4	$[50 - 50,000]$	20
			4	6	5	3	$[50 - 50,000]$	20
			5	7	3	5	$[50 - 50,000]$	20
			Avg.	**6.4**	**4.6**	**3.6**		
50	**80**	3.2	*1*	7	5	7	$[50 - 50,000]$	20
			2	7	5	5	$[50 - 50,000]$	20
			3	6	5	5	$[50 - 50,000]$	20
			4	6	5	6	$[50 - 50,000]$	20
			5	6	4	5	$[50 - 50,000]$	20
			Avg.	**6.4**	**4.8**	**5.6**		
50	**120**	4.8	*1*	10	10	7	$[50 - 50,000]$	20
			2	13	10	7	$[50 - 50,000]$	20
			3	9	6	8	$[50 - 50,000]$	20
			4	13	9	7	$[50 - 50,000]$	20
			5	12	9	7	$[50 - 50,000]$	20
			Avg.	**11.4**	**8.8**	**7.2**		
50	**160**	6.4	*1*	12	10	9	$[50 - 50,000]$	20
			2	13	9	9	$[50 - 50,000]$	20
			3	14	7	9	$[50 - 50,000]$	20
			4	11	7	8	$[50 - 50,000]$	20
			5	11	10	8	$[50 - 50,000]$	20
			Avg.	**12.2**	**8.6**	**8.6**		

TABLE B.1 **Description summary of the 30 benchmark networks used to evaluate the reconstruction methods.** The 30 benchmark networks of 50 nodes, and 20 to 160 edges, have been instantiated with the causal modeling tool Tetrad IV (http://www.phil.cmu.edu/tetrad/). For each model, 20 dataset replicates of size ranging between 50 and 50,000 were generated with Tetrad IV.

B.2 Evaluation of 3off2 by score

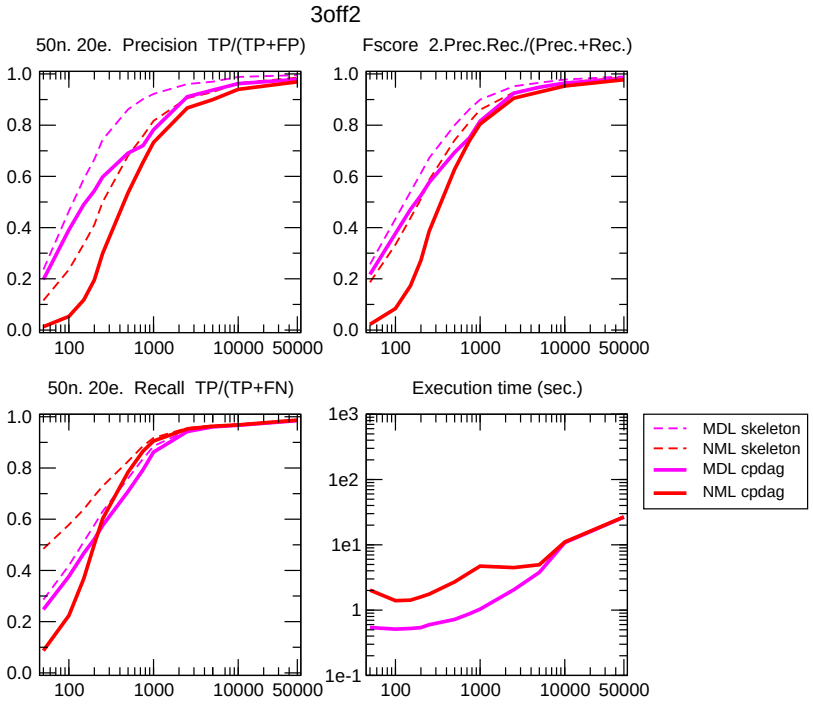


FIGURE B.1 3off2 **reconstruction, effect of complexity MDL and NML.** 50 node, 20 edge benchmark networks generated using Tetrad. $\langle k \rangle = 0.8$, $\langle k^{in}_{\max} \rangle = 2.4$ and $\langle k^{out}_{\max} \rangle = 2.2$. The change of slope in execution time at sample size $N = 1000$ for NML corresponds to the use of the Szpankowski approximation (see section 6.2.2).

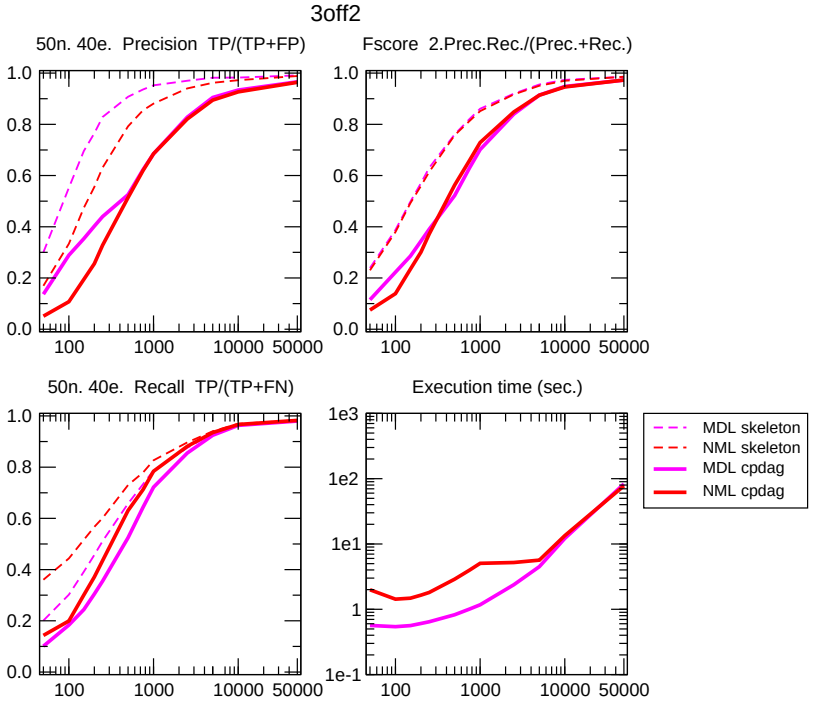


FIGURE B.2 3off2 **reconstruction, effect of complexity MDL and NML.** 50 node, 40 edge benchmark networks generated using Tetrad. $\langle k \rangle = 1.6$, $\langle k^{in}_{\max} \rangle = 3.2$ and $\langle k^{out}_{\max} \rangle = 3.6$. The change of slope in execution time at sample size $N = 1000$ for NML corresponds to the use of the Szpankowski approximation (see section 6.2.2).

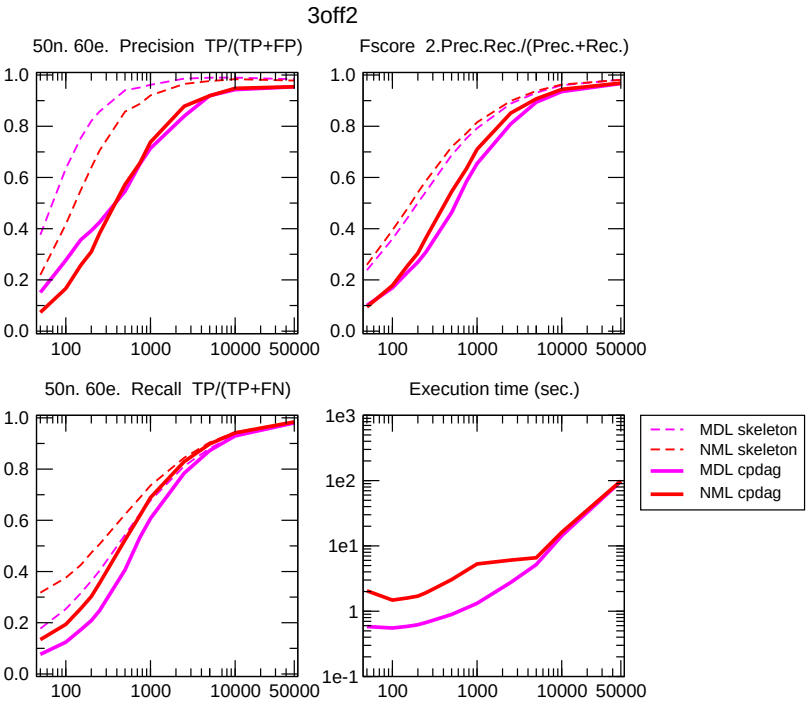


FIGURE B.3 3off2 **reconstruction, effect of complexity MDL and NML.** 50 node, 60 edge benchmark networks generated using Tetrad. $\langle k \rangle = 2.4$, $\langle k^{in}_{\max} \rangle = 4.6$ and $\langle k^{out}_{\max} \rangle = 3.6$. The change of slope in execution time at sample size $N = 1000$ for NML corresponds to the use of the Szpankowski approximation (see section 6.2.2).

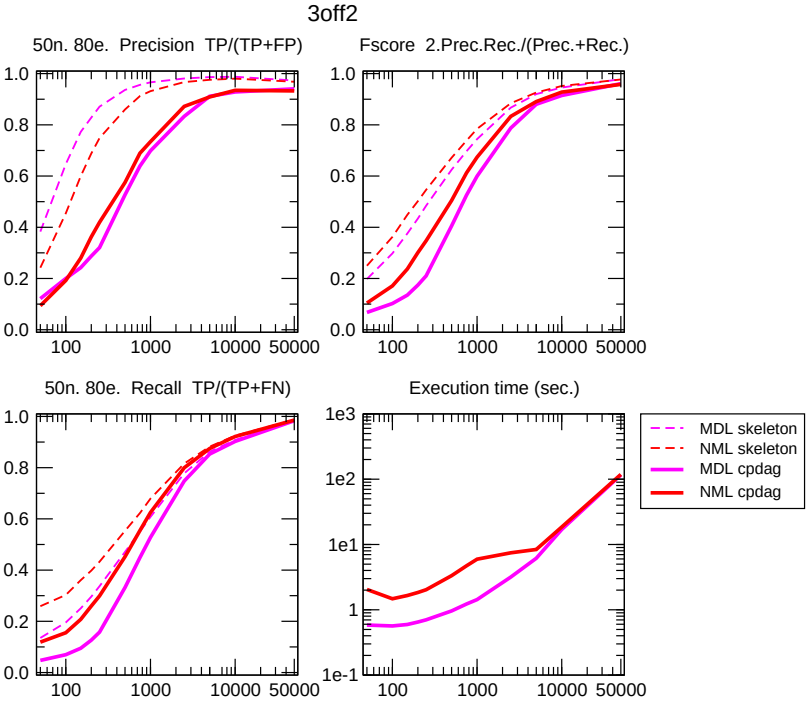


FIGURE B.4 3off2 **reconstruction, effect of complexity MDL and NML.** 50 node, 80 edge benchmark networks generated using Tetrad. $\langle k \rangle = 3.2$, $\langle k^{in}_{\max} \rangle = 4.8$ and $\langle k^{out}_{\max} \rangle = 5.6$. The change of slope in execution time at sample size $N = 1000$ for NML corresponds to the use of the Szpankowski approximation (see section 6.2.2).

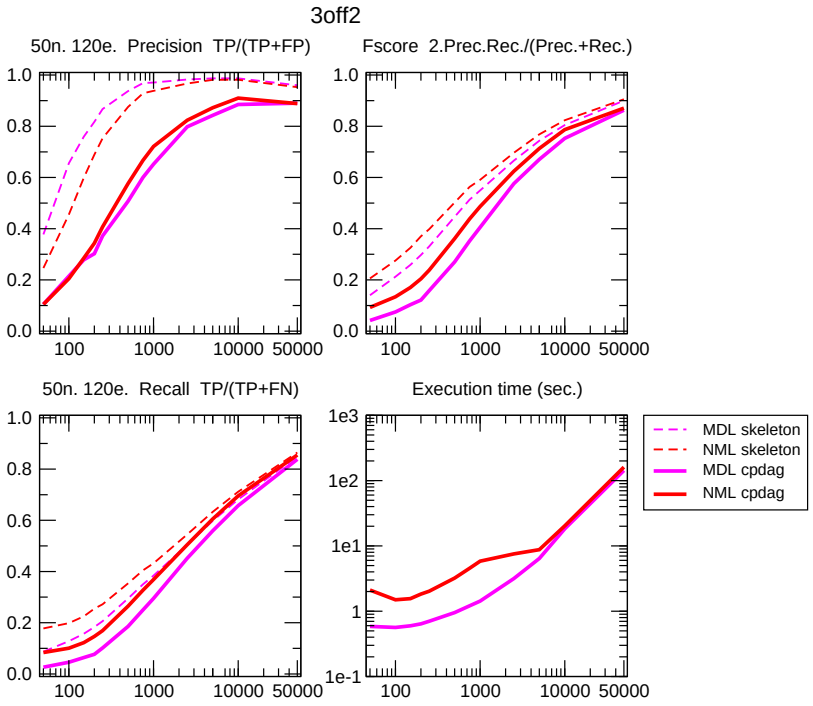


FIGURE B.5 3off2 **reconstruction, effect of complexity MDL and NML.** 50 node, 120 edge benchmark networks generated using Tetrad. $\langle k \rangle = 4.8$, $\langle k^{in}_{\max} \rangle = 8.8$ and $\langle k^{out}_{\max} \rangle = 7.2$. The change of slope in execution time at sample size $N = 1000$ for NML corresponds to the use of the Szpankowski approximation (see section 6.2.2).

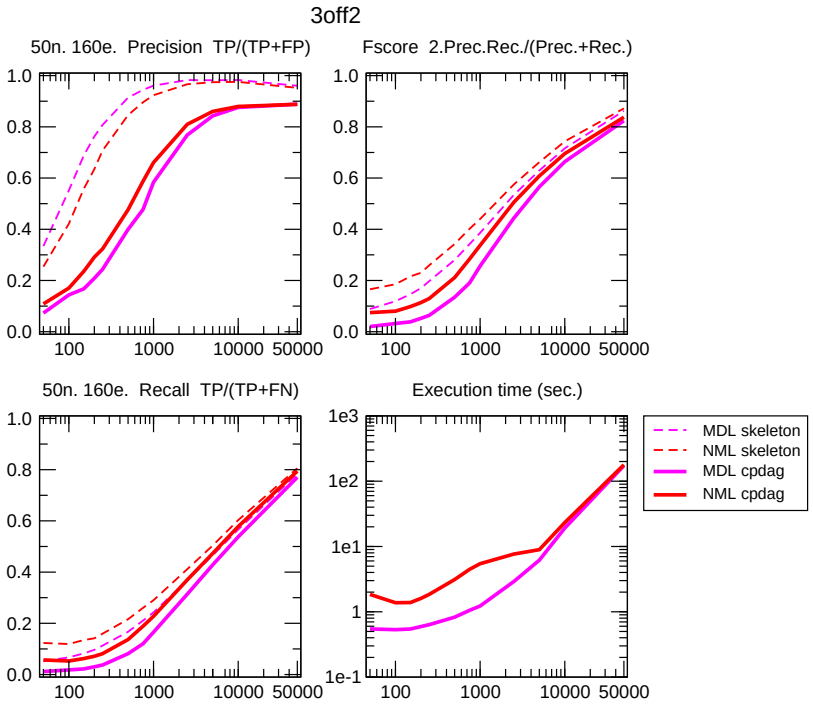


FIGURE B.6 3off2 **reconstruction, effect of complexity MDL and NML.** 50 node, 160 edge benchmark networks generated using Tetrad. $\langle k \rangle = 6.4$, $\langle k^{in}_{\max} \rangle = 8.6$ and $\langle k^{out}_{\max} \rangle = 8.6$. The change of slope in execution time at sample size $N = 1000$ for NML corresponds to the use of the Szpankowski approximation (see section 6.2.2).

B.3 Evaluation of the PC method by significance level

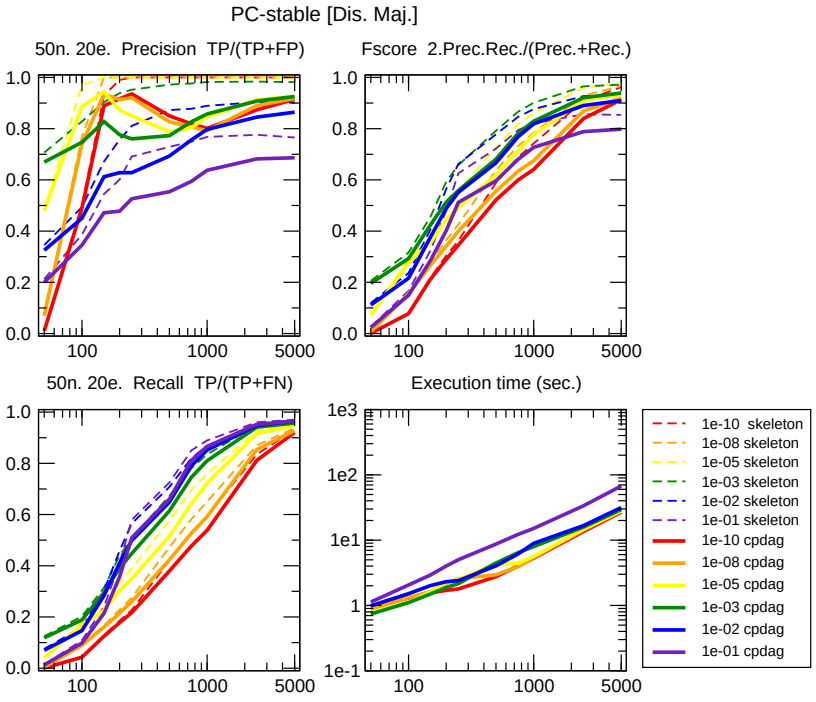


FIGURE B.7 **PC, effect of independence test parameter α.** 50 node, 20 edge benchmark networks generated using Tetrad. $\langle k \rangle = 0.8$, $\langle k^{in}_{\max} \rangle = 2.4$ and $\langle k^{out}_{\max} \rangle = 2.2$. G^2 independence test; PC-stable, majority rule [Colombo and Maathuis, 2014].

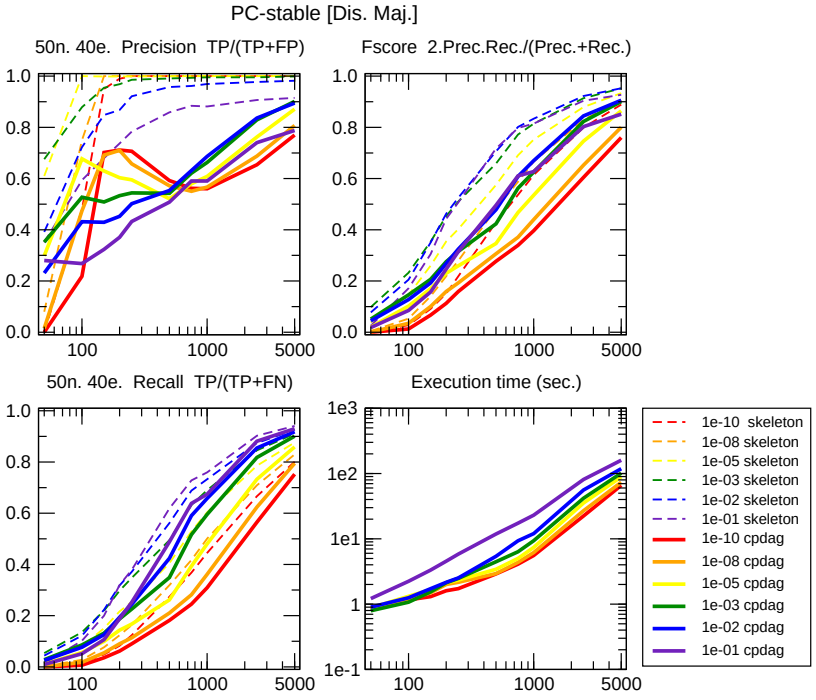


FIGURE B.8 **PC, effect of independence test parameter α.** 50 node, 40 edge benchmark networks generated using Tetrad. $\langle k \rangle = 1.6$, $\langle k^{in}_{\max} \rangle = 3.2$ and $\langle k^{out}_{\max} \rangle = 3.6$. G^2 independence test; PC-stable, majority rule [Colombo and Maathuis, 2014].

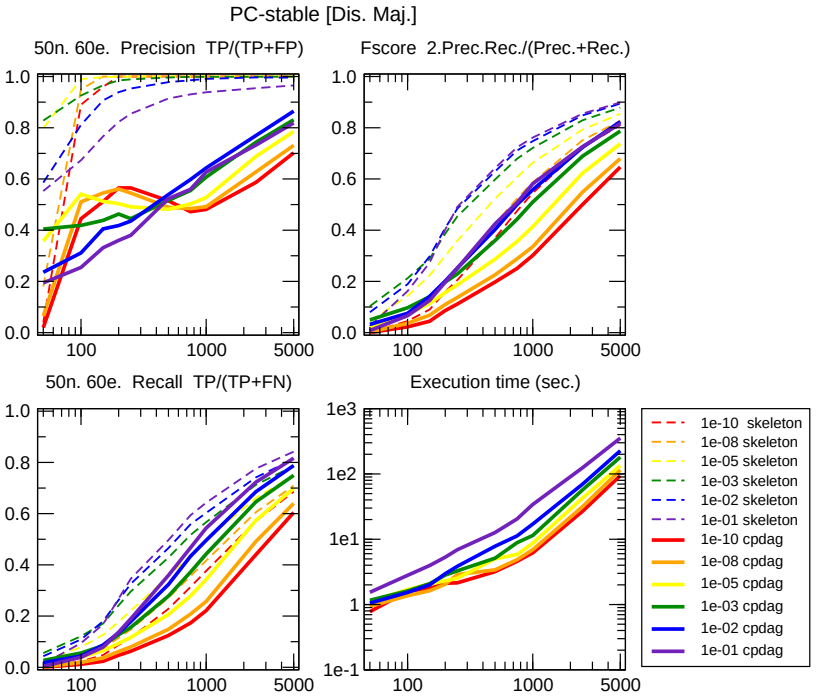


FIGURE B.9 **PC, effect of independence test parameter α.** 50 node, 60 edge benchmark networks generated using Tetrad. $\langle k \rangle = 2.4$, $\langle k^{in}_{\max} \rangle = 4.6$ and $\langle k^{out}_{\max} \rangle = 3.6$. G^2 independence test; PC-stable, majority rule [Colombo and Maathuis, 2014].

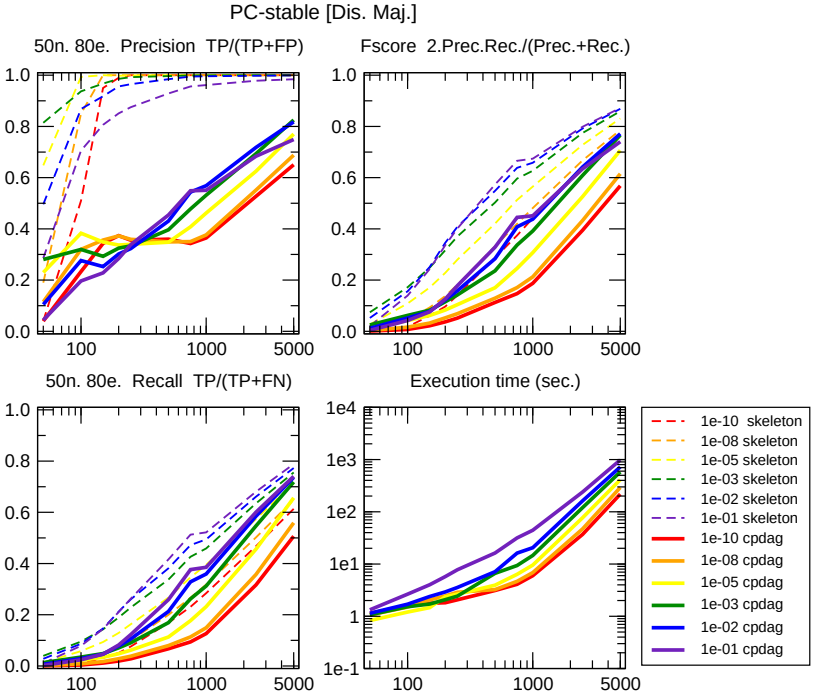


FIGURE B.10 **PC, effect of independence test parameter α.** 50 node, 80 edge benchmark networks generated using Tetrad. $\langle k \rangle = 3.2$, $\langle k^{in}_{\max} \rangle = 4.8$ and $\langle k^{out}_{\max} \rangle = 5.6$. G^2 independence test; PC-stable, majority rule [Colombo and Maathuis, 2014].

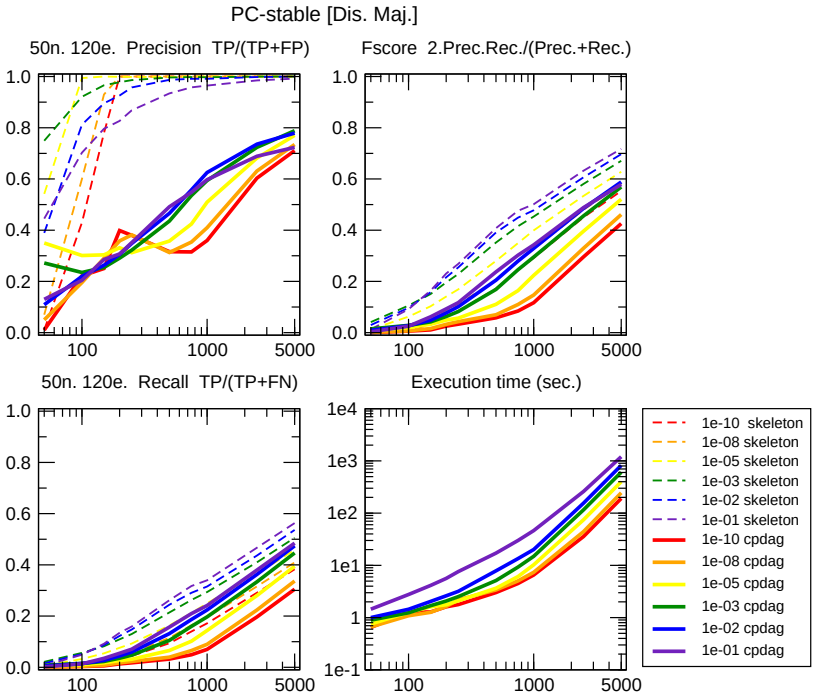


FIGURE B.11 **PC, effect of independence test parameter α.** 50 node, 120 edge benchmark networks generated using Tetrad. $\langle k \rangle = 4.8$, $\langle k^{in}_{\max} \rangle = 8.8$ and $\langle k^{out}_{\max} \rangle = 7.2$. G^2 independence test; PC-stable, majority rule [Colombo and Maathuis, 2014].

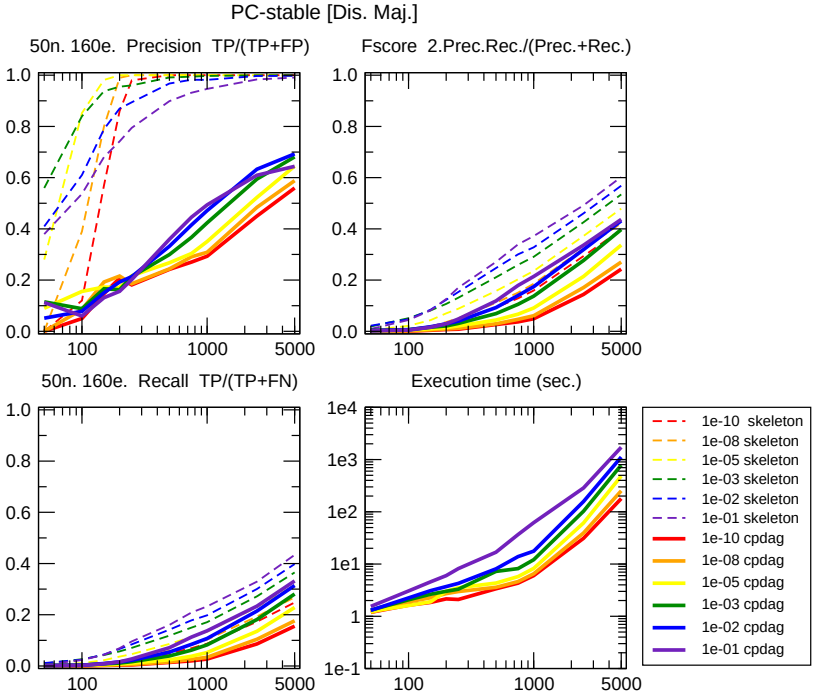


FIGURE B.12 **PC, effect of independence test parameter α.** 50 node, 160 edge benchmark networks generated using Tetrad. $\langle k \rangle = 6.4$, $\langle k^{in}_{\max} \rangle = 8.6$ and $\langle k^{out}_{\max} \rangle = 8.6$. G^2 independence test; PC-stable, majority rule [Colombo and Maathuis, 2014].

B.4 Evaluation of the MMHC method by significance level

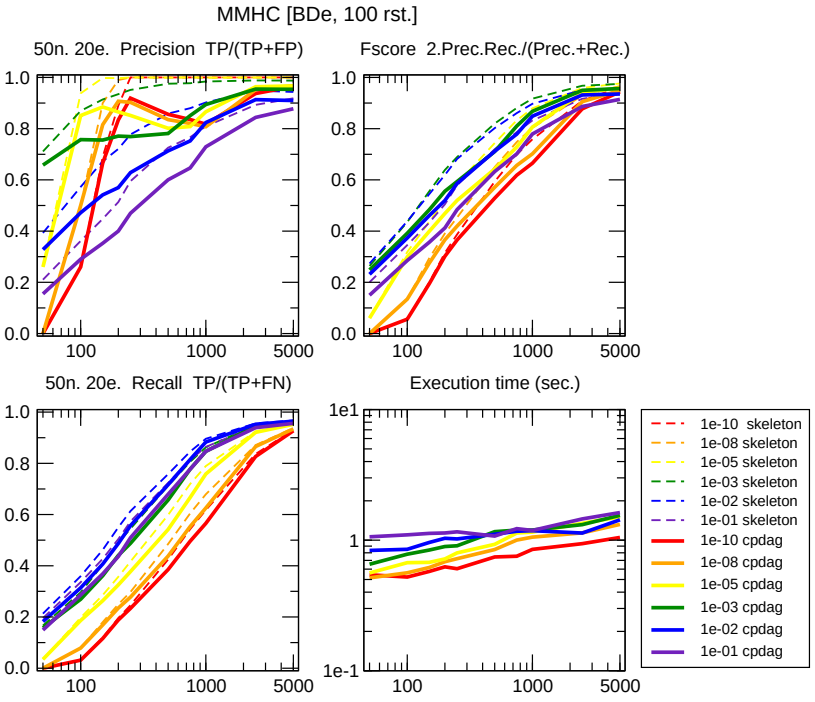


FIGURE B.13 **MMHC, effect of independence test parameter α.** 50 node, 20 edge benchmark networks generated using Tetrad. $\langle k \rangle = 0.8$, $\langle k^{in}_{\max} \rangle = 2.4$ and $\langle k^{out}_{\max} \rangle = 2.2$. G^2 independence test; MMHC, BDe score [Tsamardinos et al., 2006].

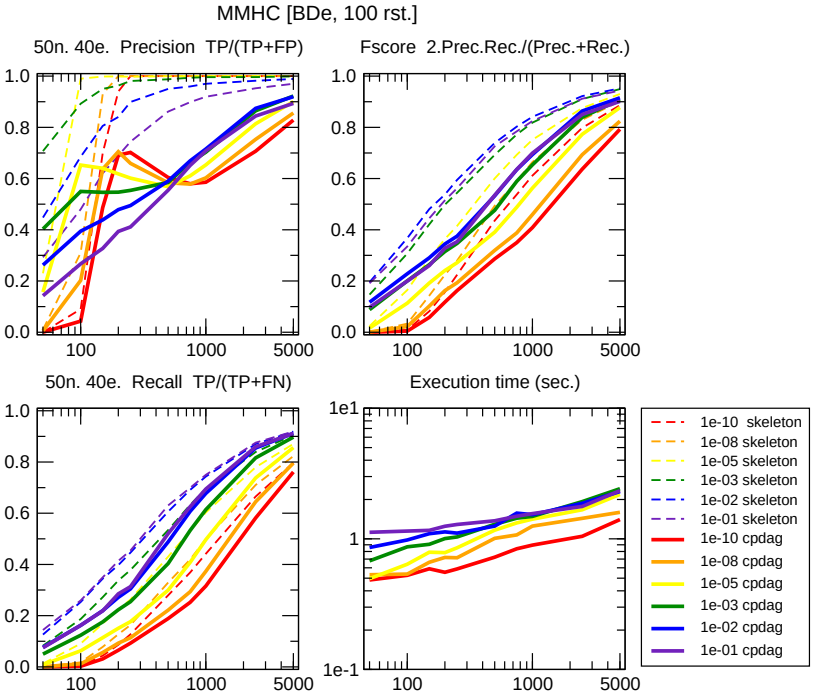


FIGURE B.14 **MMHC, effect of independence test parameter α.** 50 node, 40 edge benchmark networks generated using Tetrad. $\langle k \rangle = 1.6$, $\langle k^{in}_{\max} \rangle = 3.2$ and $\langle k^{out}_{\max} \rangle = 3.6$. G^2 independence test; MMHC, BDe score [Tsamardinos et al., 2006].

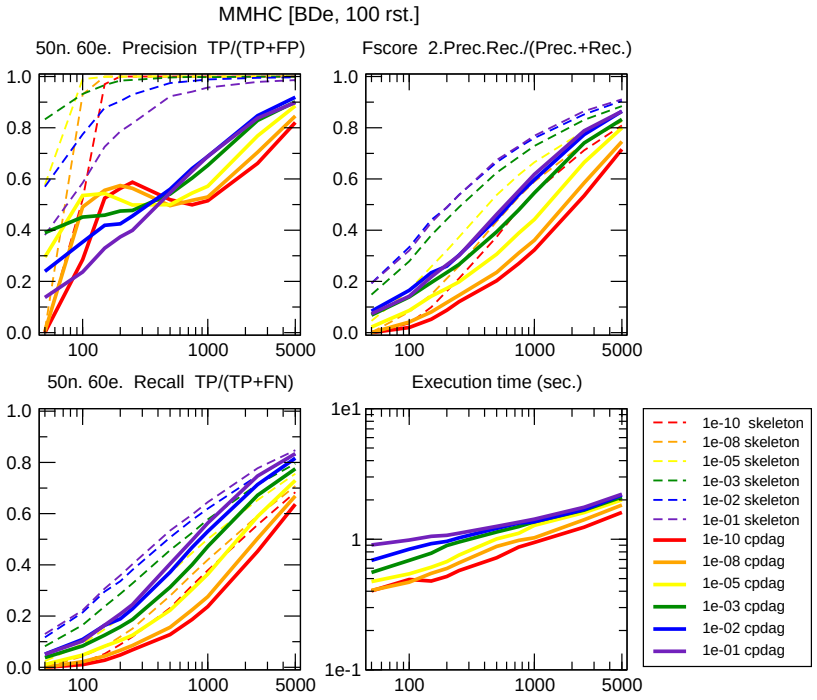


FIGURE B.15 **MMHC, effect of independence test parameter α.** 50 node, 60 edge benchmark networks generated using Tetrad. $\langle k \rangle = 2.4$, $\langle k^{in}_{\max} \rangle = 4.6$ and $\langle k^{out}_{\max} \rangle = 3.6$. G^2 independence test; MMHC, BDe score [Tsamardinos et al., 2006].

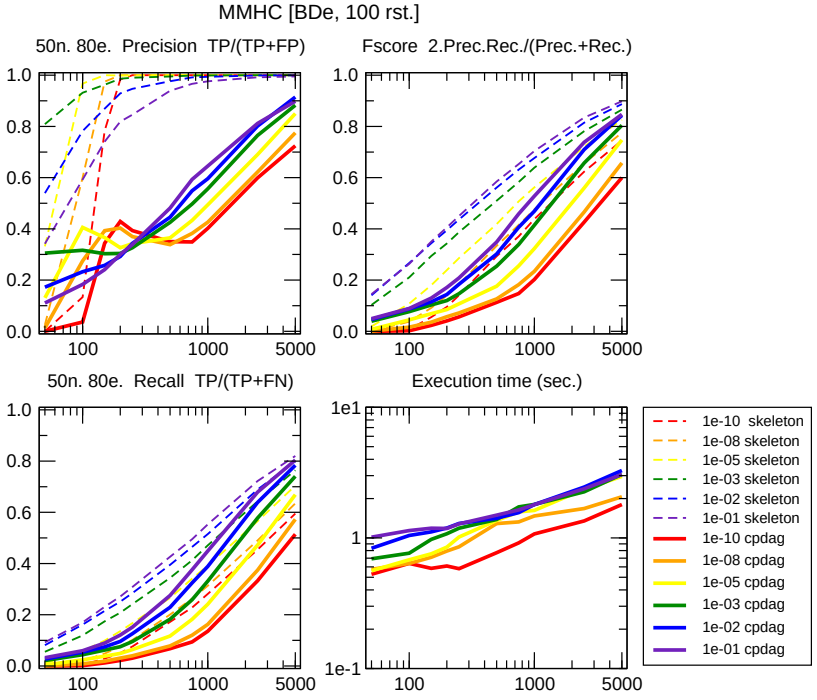


FIGURE B.16 **MMHC, effect of independence test parameter α.** 50 node, 80 edge benchmark networks generated using Tetrad. $\langle k \rangle = 3.2$, $\langle k^{in}_{\max} \rangle = 4.8$ and $\langle k^{out}_{\max} \rangle = 5.6$. G^2 independence test; MMHC, BDe score [Tsamardinos et al., 2006].

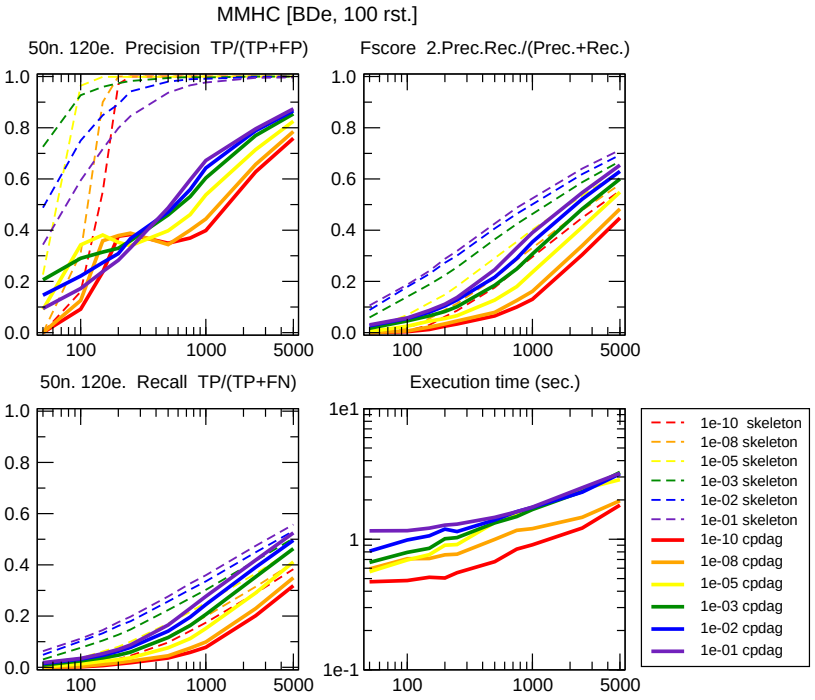


FIGURE B.17 **MMHC, effect of independence test parameter α.** 50 node, 120 edge benchmark networks generated using Tetrad. $\langle k \rangle = 4.8$, $\langle k^{in}_{\max} \rangle = 8.8$ and $\langle k^{out}_{\max} \rangle = 7.2$. G^2 independence test; MMHC, BDe score [Tsamardinos et al., 2006].

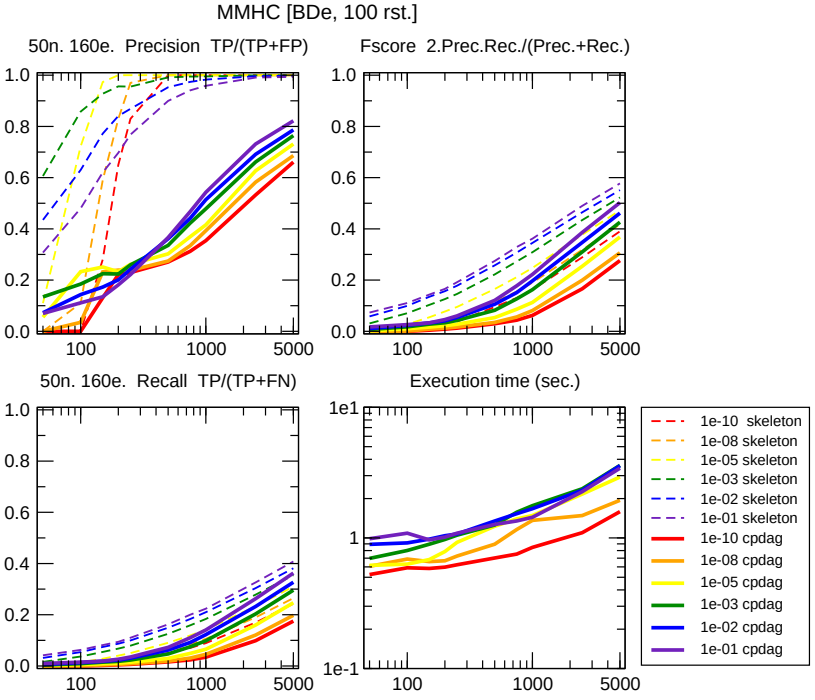


FIGURE B.18 **MMHC, effect of independence test parameter α.** 50 node, 160 edge benchmark networks generated using Tetrad. $\langle k \rangle = 6.4$, $\langle k^{in}_{\max} \rangle = 8.6$ and $\langle k^{out}_{\max} \rangle = 8.6$. G^2 independence test; MMHC, BDe score [Tsamardinos et al., 2006].

B.5 Comparison between 3off2, PC and Bayesian hill-climbing

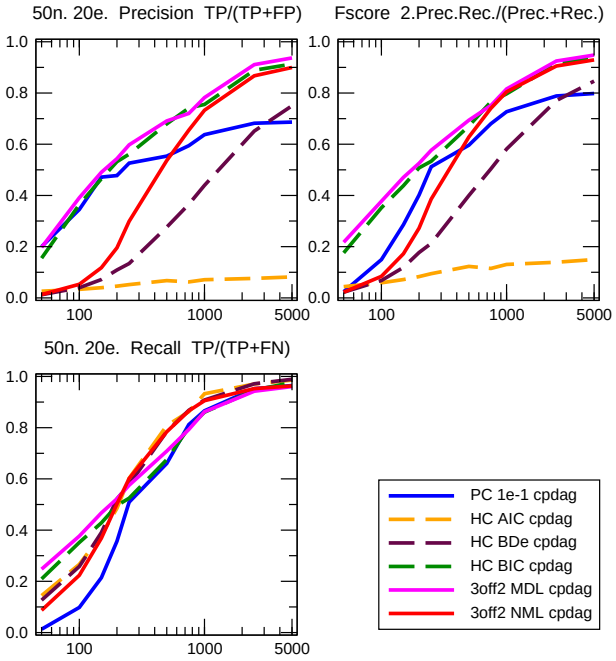


FIGURE B.19 **CPDAG comparison between 3off2, PC and Bayesian hill climbing.** 50 node, 20 edge benchmark networks generated using Tetrad. $\langle k \rangle = 0.8$, $\langle k^{in}_{\max} \rangle = 2.4$ and $\langle k^{out}_{\max} \rangle = 2.2$. Bayesian scores: AIC, BDe and BIC.

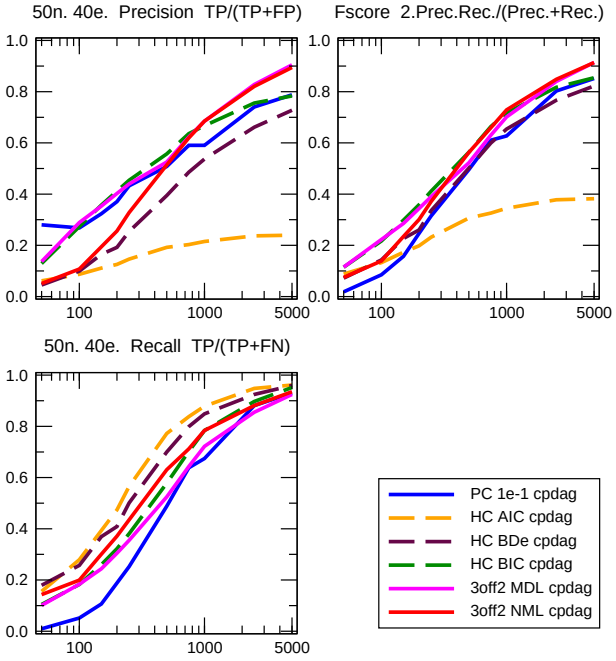


FIGURE B.20 **CPDAG comparison between 3off2, PC and Bayesian hill climbing.** 50 node, 40 edge benchmark networks generated using Tetrad. $\langle k \rangle = 1.6$, $\langle k^{in}_{\max} \rangle = 3.2$ and $\langle k^{out}_{\max} \rangle = 3.6$. Bayesian scores: AIC, BDe and BIC.

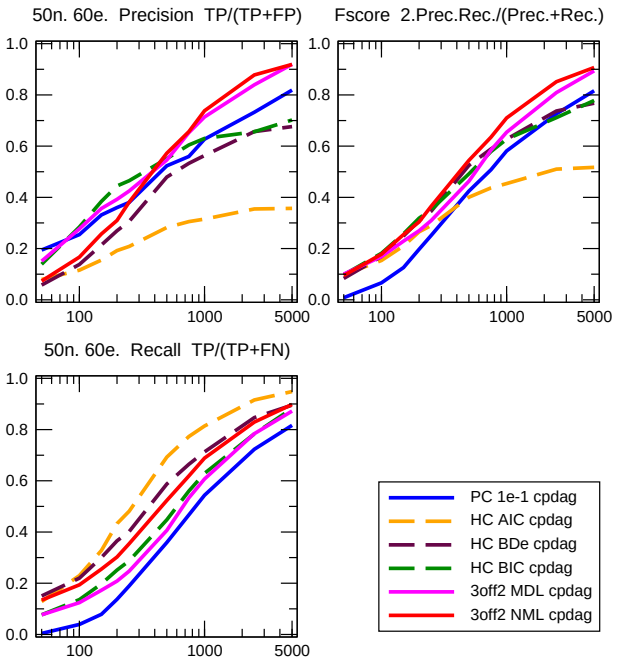


FIGURE B.21 **CPDAG comparison between 3off2, PC and Bayesian hill climbing.** 50 node, 60 edge benchmark networks generated using Tetrad. $\langle k \rangle = 2.4$, $\langle k^{in}_{\max} \rangle = 4.6$ and $\langle k^{out}_{\max} \rangle = 3.6$. Bayesian scores: AIC, BDe and BIC.

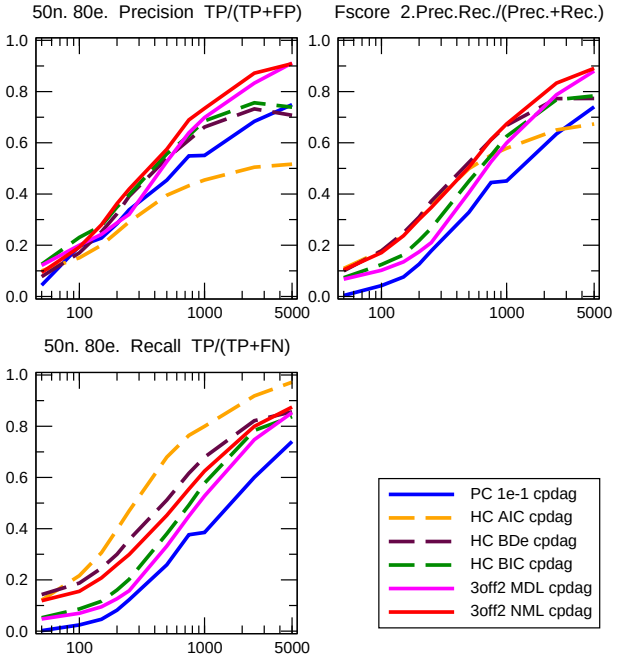


FIGURE B.22 **CPDAG comparison between 3off2, PC and Bayesian hill climbing.** 50 node, 80 edge benchmark networks generated using Tetrad. $\langle k \rangle = 3.2$, $\langle k^{in}_{\max} \rangle = 4.8$ and $\langle k^{out}_{\max} \rangle = 5.6$. Bayesian scores: AIC, BDe and BIC.

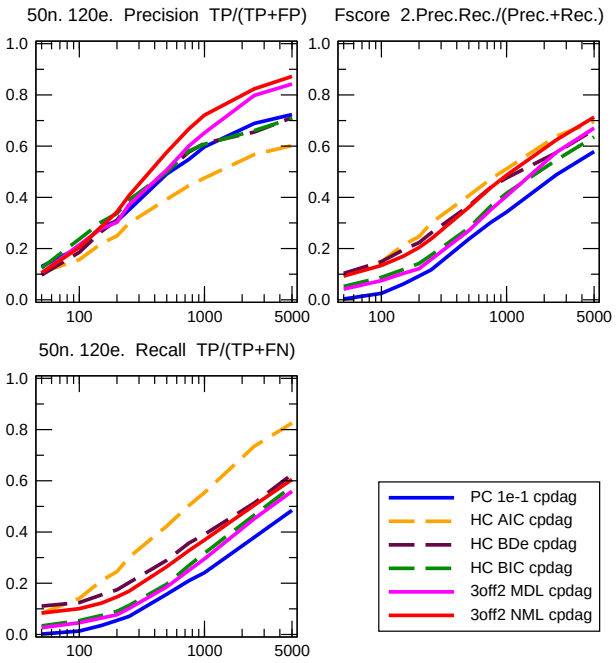


FIGURE B.23 **CPDAG comparison between 3off2, PC and Bayesian hill climbing.** 50 node, 120 edge benchmark networks generated using Tetrad. $\langle k \rangle = 4.8$, $\langle k^{in}_{\max} \rangle = 8.8$ and $\langle k^{out}_{\max} \rangle = 7.2$. Bayesian scores: AIC, BDe and BIC.

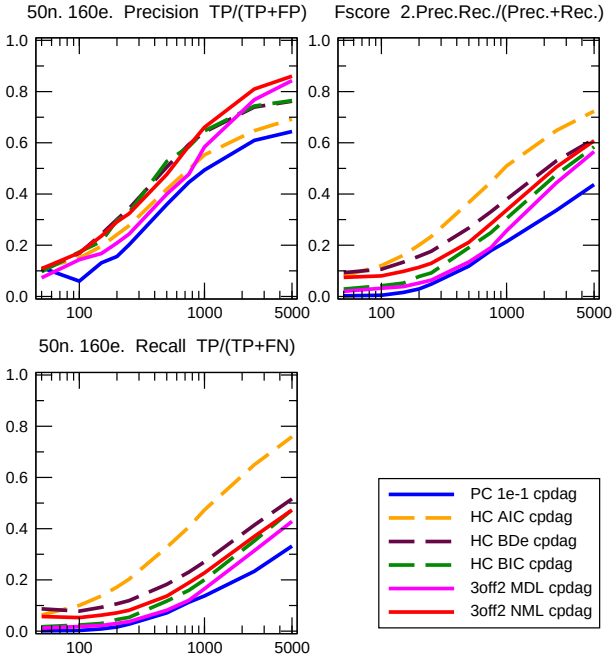


FIGURE B.24 **CPDAG comparison between 3off2, PC and Bayesian hill climbing.** 50 node, 160 edge benchmark networks generated using Tetrad. $\langle k \rangle = 6.4$, $\langle k^{in}_{\max} \rangle = 8.6$ and $\langle k^{out}_{\max} \rangle = 8.6$. Bayesian scores: AIC, BDe and BIC.

B.6 Evaluation of the Bayesian methods by score

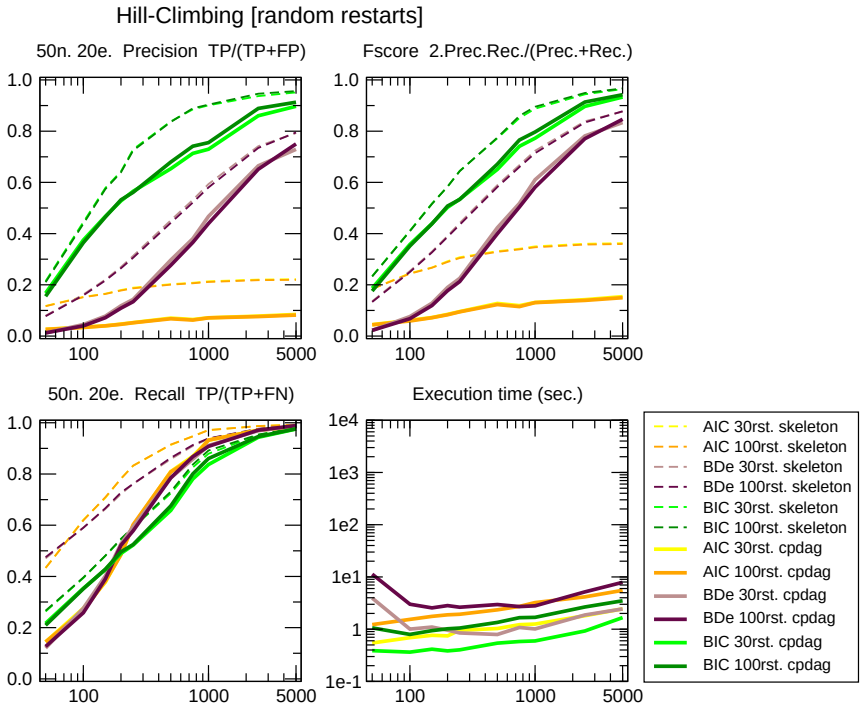


FIGURE B.25 **Bayesian Hill-Climbing, effect of Bayesian score AIC, BDe and BIC.** 50 node, 20 edge benchmark networks generated using Tetrad. $\langle k \rangle = 0.8$, $\langle k^{in}_{\max} \rangle = 2.4$ and $\langle k^{out}_{\max} \rangle = 2.2$.

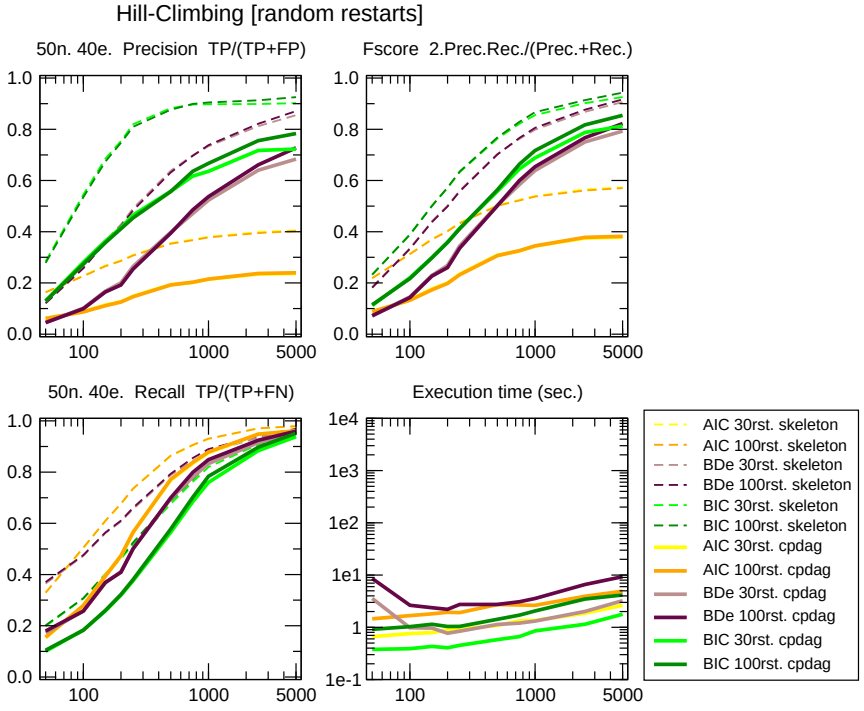


FIGURE B.26 **Bayesian Hill-Climbing, effect of Bayesian score AIC, BDe and BIC.** 50 node, 40 edge benchmark networks generated using Tetrad. $\langle k \rangle = 1.6$, $\langle k^{in}_{\max} \rangle = 3.2$ and $\langle k^{out}_{\max} \rangle = 3.6$.

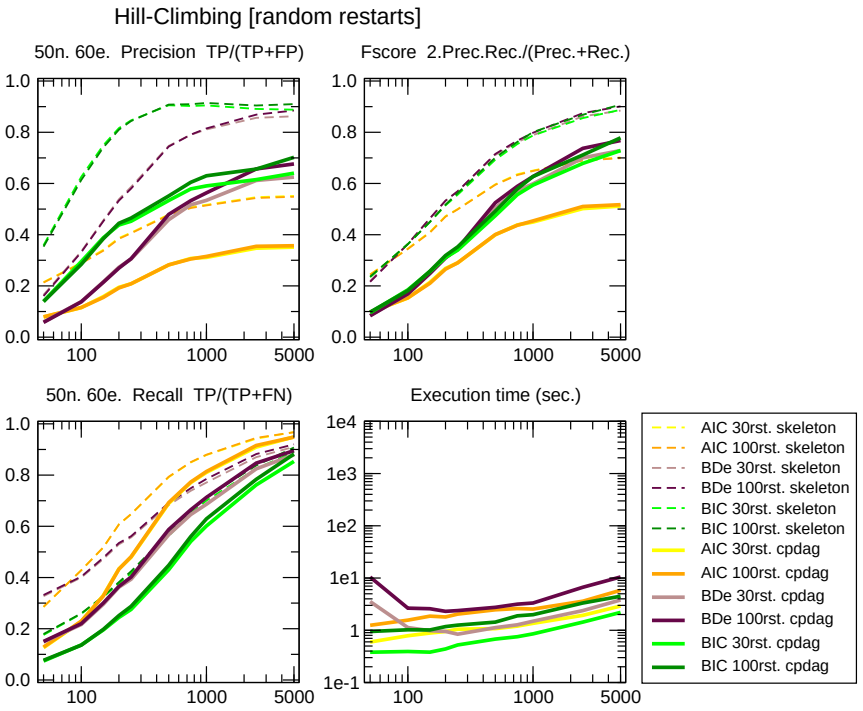


FIGURE B.27 **Bayesian Hill-Climbing, effect of Bayesian score AIC, BDe and BIC.** 50 node, 60 edge benchmark networks generated using Tetrad. $\langle k \rangle = 2.4$, $\langle k^{in}_{\max} \rangle = 4.6$ and $\langle k^{out}_{\max} \rangle = 3.6$.

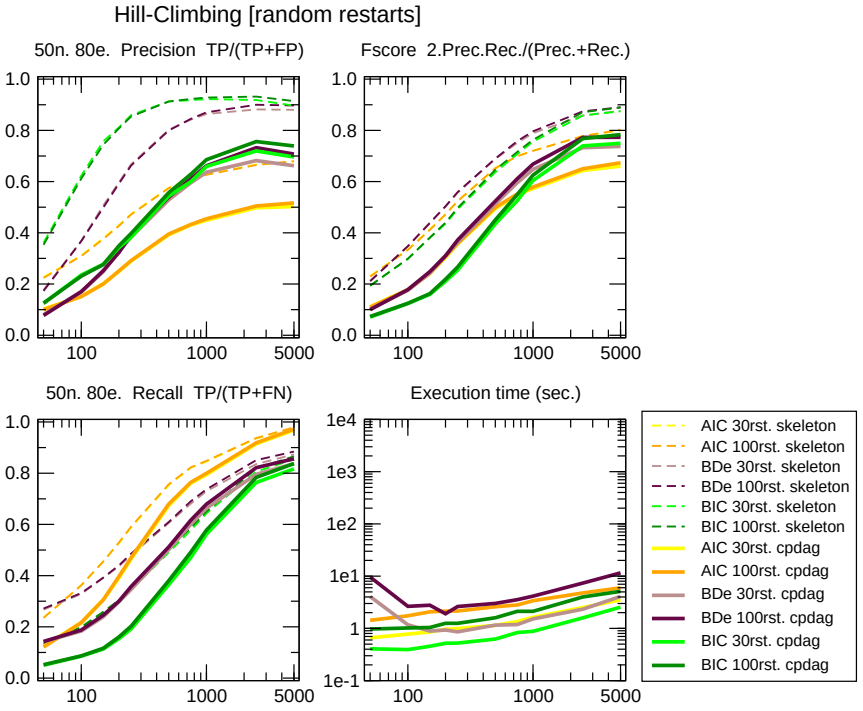


FIGURE B.28 **Bayesian Hill-Climbing, effect of Bayesian score AIC, BDe and BIC.** 50 node, 80 edge benchmark networks generated using Tetrad. $\langle k \rangle = 3.2$, $\langle k^{in}_{\max} \rangle = 4.8$ and $\langle k^{out}_{\max} \rangle = 5.6$.

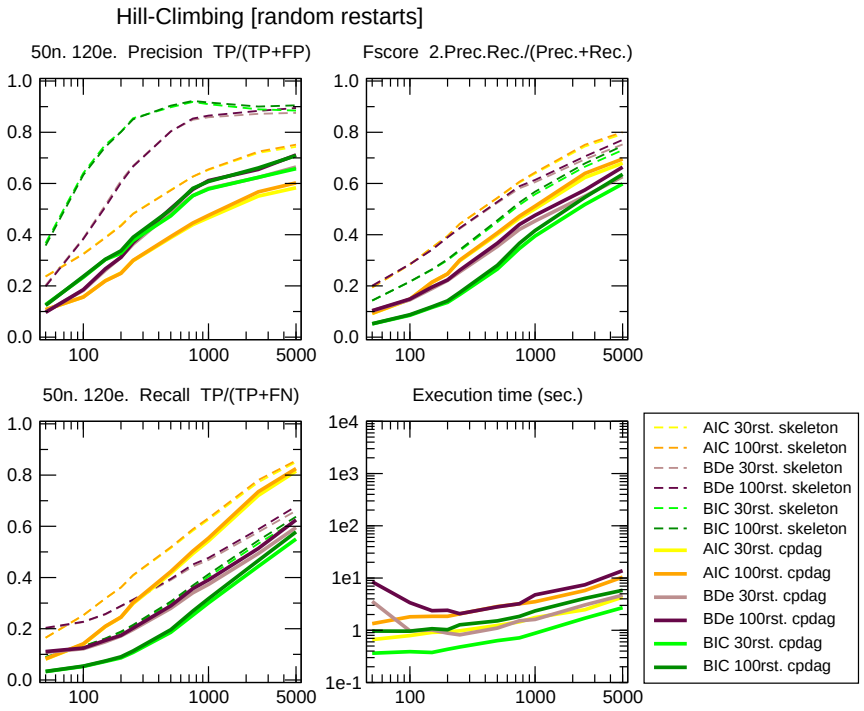


FIGURE B.29 **Bayesian Hill-Climbing, effect of Bayesian score AIC, BDe and BIC.** 50 node, 120 edge benchmark networks generated using Tetrad. $\langle k \rangle = 4.8$, $\langle k^{in}_{\max} \rangle = 8.8$ and $\langle k^{out}_{\max} \rangle = 7.2$.

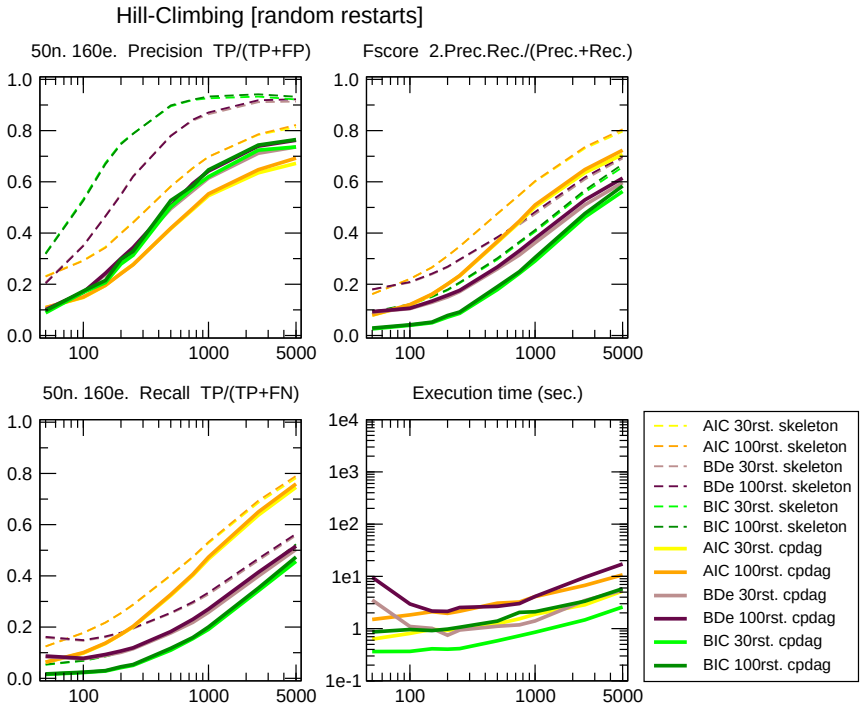


FIGURE B.30 **Bayesian Hill-Climbing, effect of Bayesian score AIC, BDe and BIC.** 50 node, 160 edge benchmark networks generated using Tetrad. $\langle k \rangle = 6.4$, $\langle k^{in}_{\max} \rangle = 8.6$ and $\langle k^{out}_{\max} \rangle = 8.6$.

Bibliography

S. Acid and L. M. de Campos. Searching for bayesian network structures in the space of restricted acyclic partially directed graphs. *J. Artif. Intell. Res.*, 18: 445–490, 2003.

S. Affeldt and H. Isambert. Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information. In *Proceedings of the 31st Conference in Uncertainty in Artificial Intelligence*, UAI'15, 2015.

S. Affeldt, P. P. Singh, I. Cascone, R. Selimoglu, J. Camonis, and H. Isambert. Évolution et cancer - expansion des familles de gènes dangereux par duplication du génome [evolution and cancer: expansion of dangerous gene repertoire by whole genome duplications] (article in french). *Med Sci (Paris)*, 29:358–361, 2013.

S. Affeldt, L. Verny, and H. Isambert. 3off2: a network reconstruction algorithm based on 2-point and 3-point information statistics. *BMC bioinformatics*, under review, 2015.

S. A. Andersson, D. Madigan, and M. D. Perlman. A characterization of markov equivalence classes for acyclic digraphs. *Ann. Statist.*, page 541, 1997.

M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo. How to infer gene networks from expression profiles. *Molecular systems biology*, 3, February 2007.

G. Berx, A. M. Cleton-Jansen, F. Nollet, W. J. de Leeuw, M. van de Vijver, C. Cornelisse, and F. van Roy. E-cadherin is a tumour/invasion suppressor gene mutated in human lobular breast cancers. *EMBO J.*, 14:6107–6115, 1995.

I.H. Bianco and F. Engert. Visuomotor transformations underlying hunting behavior in zebrafish. *Current Biology*, 25:831–846, 2015.

I.H. Bianco and S.W. Wilson. Whole-brain activity maps reveal stereotyped, distributed networks for visuomotor behavior. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 364:1005–20, 2009.

G. Bontempi and P. E. Meyer. Causal filter selection in microarray data. In *ICML*, pages 95–102, 2010.

R. R. Bouckaert. Probabilistic network construction using the minimum description length principle. *In Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, 747:41–48, 1993.

R. R. Bouckaert. Properties of bayesian belief network learning algorithms. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*, UAI'94, pages 102–109, 1994.

W. Buntine. Theory refinement on bayesian networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, pages 52–60, 1991.

A. J. Butte and I. S. Kohane. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, 5:415–426, 2000.

A. Cano, M. Gomez-Olmedo, and S. Moral. A score based ranking of the edges for the pc algorithm. In *Proceedings of the European Workshop on Probabilistic Graphical Models (PGM)*, pages 41–48, 2008.

I. Cantone, L. Marucci, F. Iorio, M. A. Ricci, V. Belcastro, M. Bansal, S. Santini, M. di Bernardo, D. di Bernardo, and M. P. Cosma. A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches. *Cell*, 137:172–181, 2009.

W. Y. I. Chan, G. A. Follows, G. Lacaud, J. E. Pimanda, J.-R. Landry, S. Kinston, K. Knezevic, S. Piltz, I. J. Donaldson, L. Gambardella, F. Sablitzky, A. R. Green, V. Kouskoff, and B. Göttgens. The paralogous hematopoietic regulators lyl1 and scl are coregulated by ets and gata factors, but lyl1 cannot rescue the early scl–/– phenotype. *Blood*, 109(5):1908–1916, 2006.

L. W. Cheung, B. T. Hennessy, J. Li, S. Yu, A. P. Myers, B. Djordjevic, Y. Lu, K. Stemke-Hale, M. D. Dyer, F. Zhang, Z. Ju, L. C. Cantley, S. E. Scherer, H. Liang, K. H. Lu, R. R. Broaddus, and G. B. Mills. High frequency of PIK3R1 and PIK3R2 mutations in endometrial cancer elucidates a novel mechanism for regulation of pten protein stability. *Cancer Discov.*, 1:180–85, 2011.

D. M. Chickering. Learning equivalence classes of bayesian-network structures. *Journal of Machine Learning Research*, 2:445–498, 2002.

D. M. Chickering, D. Geiger, and D. Heckerman. Learning Bayesian networks: Search methods and experimental results. In *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, pages 112–128, 1995.

C. K. Chow and C. N. Liu. Approximating Discrete Probability Distributions With Dependence Trees. *IEEE Transactions on Information Theory*, IT-14: 462–467, 1968.

A. H. Chowdhury, J. R. Ramroop, G. Upadhyay, A. Sengupta, A. Andrzejczyk, and S. Saleque. Differential transcriptional regulation of meis1 by gfi1b and its co-factors lsd1 and corest. *PLoS ONE*, 8(1), 01 2013.

M. Cizkova, S. Vacher, D. Meseure, M. Trassard, A. Susini, D. Mlcuchova, C. Callens, E. Rouleau, F. Spyratos, R. Lidereau, and I. Bièche. Pik3r1 underexpression is an independent prognostic marker in breast cancer. *BMC Cancer*, 13: 1471–2407, 2013.

T. Claassen and T. Heskes. A bayesian approach to constraint based causal inference. In *In Proc. of the 28th Conference on Uncertainty in Artificial Intelligence (UAI'12)*, pages 207–216, 2012.

S. M. Cleveland, S. Smith, R. Tripathi, E. M. Mathias, C. Goodings, N. Elliott, D. Peng, W. El-Rifai, D. Yi, X. Chen, L. Li, C. Mullighan, J. R. Downing, P. Love, and U. P. Davé. Lmo2 induces hematopoietic stem cell like features in T-cell progenitor cells prior to leukemia. *Stem Cells*, 31(4):882 – 894, 2013.

D. Colombo and M. H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15:3741–3782, 2014.

D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, 02 2012.

G. F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Mach. Learn.*, 9(4):309–347, 1992.

T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

D. Dash and M. J. Druzdzel. A hybrid anytime algorithm for the construction of causal models from sparse data. In *Proceedings of the Fifteenth International Conference on Uncertainty in Artificial Intelligence*, pages 142–149, 1999.

L. M. de Campos. A scoring function for learning bayesian networks based on mutual information and conditional independence tests. *J. Mach. Learn. Res.*, 7:2149–2187, 2006.

L. M. de Campos and J. F. Huete. Technologies for constructing intelligent systems. chapter Stochastic Algorithms for Searching Causal Orderings in Bayesian Networks, pages 327–340. 2002.

N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, W. Schierding, D. C. Koboldt, T. B. Mooney, M. B. Callaway, D. Dooling, E. R. Mardis, R. K. Wilson, and L. Ding. MuSiC: identifying mutational significance in cancer genomes. *Genome research*, 22(8):1589–1598, 2012.

J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, J. E. Richardson, M. Airey, A. Anagnostopoulos, R. Babiuk, R. Baldarelli, et al. The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res.*, 40:D881–886, 2012.

J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, 5:8, 2007.

S. A. Forbes, G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J. W. Teague, P. A. Futreal, and M. R. Stratton. The catalogue of somatic mutations in cancer (COSMIC). *Current Protocols in Human Genetics*, 2008.

A. M. Fraser and H. L. Swinney. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A.*, 33:1134–1140, 1986.

N. Friedman. The bayesian structural em algorithm. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pages 129–138, 1998.

N. Friedman and D. Koller. Being bayesian about network structure. A bayesian approach to structure discovery in bayesian networks. *Machine Learning*, 50 (1-2):95–125, 2003.

N. Friedman, I. Nachman, and D. Peér. Learning bayesian network structure from massive datasets: The 'sparse candidate' algorithm. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, pages 206–215, 1999.

M. Gasco, S. Shami, and T. Crook. The p53 pathway in breast cancer. *Breast Cancer Res.*, 4:70–76, 2002.

Berthold Göttgens, Aristotelis Nastos, Sarah Kinston, Sandie Piltz, Eric C.M. Delabesse, Maureen Stanley, Maria-Jose Sanchez, Aldo Ciau-Uitz, Roger Patient, and Anthony R. Green. Establishing the transcriptional programme for blood: the scl stem cell enhancer is regulated by a multiprotein complex containing ets and gata factors. *The EMBO Journal*, 21(12):3039–3050, 2002.

J. A. Grass, M. E. Boyer, S. Pal, J. Wu, M. J. Weiss, and E. H. Bresnick. GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. *Proc. Natl. Acad. Sci. USA*, 100(15):8811 – 8816, 07 2003.

T. S. Han. Multiple mutual informations and multiple interactions in frequency data. *Information and Control*, 46(1):26–45, 1980.

M. H. Hansen and B. Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96:746–774, 2001.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, 20(3): 197–243, September 1995. Available as Technical Report MSR-TR-94-09.

M. E. Higgins, M. Claremont, J.E. Major, C. Sander, and A.E. Lash. CancerGenes: a gene selection resource for cancer genome projects. *Nucleic acids research*, 35 (suppl 1):D721, 2006.

T. F. Holekamp, D. Turaga, and T. E. Holy. Fast three-dimensional fluorescence imaging of activity in neural populations by objective-coupled planar illumination microscopy. *Neuron*, 57(5):661 – 672, 2008.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *J. Mach. Learn. Res.*, 8:613–636, May 2007.

M. Kalisch and P. Bühlmann. Robustification of the pc-algorithm for directed acyclic graphs. *Journal Of Computational And Graphical Statistics*, 17(4):773–789, 2008.

M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann. Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

P. J. Keller, A. D. Schmidt, A. Santella, K. Khairy, Z. Bao, J. Wittbrodt, and E. H. K. Stelzer. Fast, high-contrast imaging of animal development with scanned light sheet-based structured-illumination microscopy. *Nat. Meth.*, 7 (8):637–642, 2010.

M. Khandekar, W. Brandt, S. Dagenais, TW. Glover, N. Suzuki, R. Shimizu, M. Yamamoto, K. C. Lim, and J. D. Engel. A gata2 intronic enhancer confers its pan-endothelia-specific regulation. *Development*, 134(9):1703 – 1712, 2007.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

T. Kocka and R. Castelo. Improved learning of bayesian networks. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pages 269–276, 2001.

M. Koivisto and K. Sood. Exact bayesian structure discovery in bayesian networks. *J. Mach. Learn. Res.*, 5:549–573, 2004.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

P. Kontkanen. Computationally efficient methods for mdl-optimal density estimation and data clustering. *Ph.D. Dissertation*, 2009.

P. Kontkanen and P. Myllymäki. A linear-time algorithm for computing the multinomial stochastic complexity. *Inf. Process. Lett.*, 103(6):227–233, 2007.

P. Kontkanen, W. Buntine, P. Myllymäki, J. Rissanen, and H. Tirri. Efficient computation of stochastic complexity. *in: C. Bishop, B. Frey (Eds.) Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics, Society for Artificial Intelligence and Statistics*, 103:233–238, 2003.

F. Kubo, B. Hablitzel, M. Dal Maschio, W. Driever, H. Baier, and A.B. Arrenberg. Functional architecture of an optic flow-responsive area that drives horizontal eye movements in zebrafish. *Neuron*, 81, 2014.

P. Larrañaga, C. M. H. Kuijpers, R. H. Murga, and Y. Yurramendi. Learning bayesian network structures by searching for the best ordering with genetic

algorithms. *IEEE Transactions on Systems, Man and Cybernetics*, 26:487–493, 1996.

M. S. et al Lawrence. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, pages 1–5, jun 2013.

H. Lei, S. Sjöberg-Margolin, S. Salahshor, B. Werelius, E. Jandáková, K. Hemminki, A. Lindblom, and I. Vorechovský. CDH1 mutations are present in both ductal and lobular breast cancer, but promoter allelic variants show no detectable breast cancer risk. *Int. J. Cancer.*, 98:199–204, 2002.

L. Li, R. Jothi, K. Cui, J. Y. Lee, T. Cohen, M. Gorivodsky, I. Tzchori, Y. Zhao, S. M. Hayes, E. H. Bresnick, K. Zhao, H. Westphal, and P. E. Love. Nuclear adaptor ldb1 regulates a transcriptional program essential for the maintenance of hematopoietic stem cells. *Nat. Immunol.*, pages 129 – 136, 02 2011.

K-C Liang and X. Wang. Gene regulatory network reconstruction using conditional mutual information. *EURASIP J. Bioinformatics and Systems Biology*, 2008.

T. Makino and A. McLysaght. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc. Natl. Acad. Sci. USA*, 107(20): 9270, 2010.

G. Malaguti, P. P. Singh, and H. Isambert. On the retention of gene duplicates prone to dominant deleterious mutations. *Theor Popul Biol.*, 93:38–51, 2014.

B. Malone, C. Yuan, and E. Hansen. Memory-efficient dynamic programming for learning optimal bayesian networks. 2011.

A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera, and A. Califano. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006.

W. J. McGill. Multivariate information transmission. *Trans. of the IRE Professional Group on Information Theory (TIT)*, 4:93–111, 1954.

C. Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence,* Montreal, QU, pages 403–418. Morgan Kaufmann, , August 1995.

C. Mermel, S. Schumacher, B. Hill, M. Meyerson, R. Beroukhim, and G. Getz. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, 12(4): R41+, 2011.

P. Meyer, D. Marbach, S. Roy, and M. Kellis. Information-theoretic inference of gene networks using backward elimination. In *BIOCOMP*, pages 700–705. CSREA Press, 2010.

P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J. Bioinformatics and Systems Biology*, 2007, 2007.

P. E. Meyer, F. Lafitte, and G. Bontempi. minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, 9:461, 2008.

Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., 1 edition, 1997.

V. Moignard, I. C. Macaulay, G. Swiers, F. Buettner, J. Schütte, F. J. Calero-Nieto, S. Kinston, A. Joshi, R. Hannah, F. J. Theis, S. E. Jacobsen, M. F. de Bruijn, and B. Göttgens. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat Cell Biol.*, 15:363 – 372, 2013.

P. Naïm, P.-H. Wuillemin, P. Leray, O. Pourret, and A. Becker. *Réseaux bayésiens*. Algorithmes. Eyrolles, 3e éd, Paris, 2011.

S. Ohno. *Evolution by gene duplication.* London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag., 1970.

S. Ohno, U. Wolf, and N.B. Atkin. Evolution from fish to mammals by gene duplication. *Hereditas*, 59(1):169–187, 1968.

S. H. Oram, J.A.I. Thoms, C. Pridans, M.E. Janes, S.J. Kinston, S. Anand, J.R. Landry, R. B. Lock, P-S. Jayaraman, B.J. Huntly, J. E. Pimanda, and B. Göttgens. A previously unrecognized promoter of lmo2 forms part of a transcriptional regulatory circuit mediating lmo2 expression in a subset of t-acute lymphoblastic leukaemia patients. *Oncogene*, pages 5796 – 5808, 10 2010.

T. Panier, S. A. Romano, R. Olive, T. Pietri, G. Sumbre, R. Candelier, and G. Debrégeas. Fast functional imaging of multiple brain regions in intact zebrafish larvae using selective plane illumination microscopy. *Frontiers in Neural Circuits*, 7, 2013.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

J. Pearl. Direct and Indirect Effects. *In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, San Francisco, CA: Morgan Kaufmann*, pages 411–420, June 2001.

J. Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, 2nd edition, 2009.

J. Pearl. The causal mediation formula–a guide to the assessment of pathways and mechanisms. *Prevention Science*, 13(4):426–436, Aug 2012.

J. Pearl and T. Verma. A theory of inferred causation. In *In Knowledge Representation and Reasoning: Proc. of the Second Int. Conf.*, pages 441–452. , 1991.

H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8):1226–1238, 2005.

R. Portugues, C. E. Feierstein, F. Engert, and M. B. Orger. Whole-brain activity maps reveal stereotyped, distributed networks for visuomotor behavior. *Neuron*, 81:875–81+, 2014.

N. H. Putnam, T. Butts, D. E. K. Ferrier, R. F. Furlong, U. Hellsten, T. Kawashima, M. Robinson-Rechavi, E. Shoguchi, A. T. Jr-Kai Yu, et al. The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453 (7198):1064–1071, 2008.

P. Ramdya, B. Reiter, and F. Engert. Reverse correlation of rapid calcium signals in the zebrafish optic tectum in vivo. *J. Neurosci. Methods.*, 157, 2006.

J. Ramsey, J. Zhang, and P. Spirtes. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 401–408, 2006.

G. Rebane and J. Pearl. The recovery of causal poly-trees from statistical data. *Int. J. Approx. Reasoning*, 2(3):341, 1988.

H. Reichenbach. *The Direction of Time*. University of California Press, 1956.

T. Richardson and P. Spirtes. Ancestral graph markov models. *Ann. Statist.*, 30: 962–1030, 2002.

T. S. Richardson. . In *Learning in Graphical Models*, chapter Chain Graphs and Symmetric Associations, pages 231–259. MIT Press, 1999.

J. Rissanen. Modeling by shortest data description. *Automatica*, vol. 14:465–471, 1978.

J. Rissanen and I. Tabus. Kolmogorov's structure function in mdl theory and lossy data compression. In *Adv. Min. Descrip. Length Theory Appl.*, page 10. MIT Press, , 2005.

T. J. W. Robinson, J. C. Liu, F. Vizeacoumar, T Sun, N. Maclean, S. E. Egan, A. D. Schimmer, A. Datti, and E. Zacksenhaus. RB1 status in triple negative breast cancer cells dictates response to radiation treatment and selective therapeutic drugs. *PLoS ONE*, 8(11), 11 2013.

T. Roos, T. Silander, P. Kontkanen, and P. Myllymäki. Bayesian network structure learning using factorized nml universal models. In *Proc. 2008 Information Theory and Applications Workshop (ITA-2008)*, , 2008a. IEEE Press. invited paper.

T. Roos, T. Silander, P. Kontkanen, and P. Myllymäki. Bayesian network structure learning using factorized nml universal models, 2008b.

D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–590, 1976.

K. Sachs, O. Perez, D. Pe'er, D.A. Lauffenburger, and G.P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308 (5721):523, 2005.

I. N. Sanov. On the probability of large deviations of random variables. *Mat. Sbornik*, 42:11–44, 1957.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6: 461–464, 1978.

M. Scutari. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, 35(3):1–22, 2010.

C. Shannon. A mathematical theory of communication. *Bell Systems Techn. Journal*, 27:623–656, 1948.

Y. M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission (Translated from)*, 23(3):3–17, 1987.

T. Silander and P. Myllymaki. A simple approach for finding the globally optimal bayesian network structure. In *Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 445–452, Arlington, Virginia, 2006. AUAI Press.

M. Singh and M. Valtorta. An algorithm for the construction of bayesian network structures from data. In *International Journal of Approximate Reasoning*, pages 259–265, 1993.

P. P. Singh, S. Affeldt, I. Cascone, R. Selimoglu, J. Camonis, and H. Isambert. On the expansion of 'dangerous' gene repertoires by whole-genome duplications in early vertebrates. *Cell Reports*, 2012.

P. P. Singh, S. Affeldt, G. Malaguti, and H. Isambert. Human dominant disease genes are enriched in paralogs originating from whole genome duplication. *PLoS Comput Biol.*, 10, 2014.

P. Spirtes. An anytime algorithm for causal inference. In *in the Presence of Latent Variables and Selection Bias in Computation, Causation and Discovery*, pages 121–128. MIT Press, 2001.

P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9:62–72, 1991.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer Verlag, New York, 1993.

P. Spirtes, C. Meek, and T. Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. In *Computation, Causation, and Discovery*, pages 211–252. AAAI Press, 1999.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

C. J. Spooner, J. X. Cheng, E. Pujadas, P. Laslo, and H. Singh. A recurrent network involving the transcription factors pu.1 and gfi1 orchestrates innate and adaptive immune cell fates. *Immunity*, 31(4):576 – 586, 2009. ISSN 1074-7613.

P. J. et al. Stephens. The landscape of cancer genes and mutational processes in breast cancer. *Nature*, page 400–404, 2012.

R. Steuer, J. Kurths, C.O. Daub, J. Weise, and J. Selbig. The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, 18 Suppl.2:S231–S240, 2002.

W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. John Wiley & Sons, Inc., 2001.

TCGA. Comprehensive molecular portraits of human breast tumours. *Nature*, 490:61–70, 2012.

L. Tian, S. A. Hires, T. Mao, D. Huber, M. E. Chiappe, S. H. Chalasani, L. Petreanu, J. Akerboom, S. A. McKinney, E. R. Schreiter, C. I. Bargmann, V. Jayaraman, K. Svoboda, and L. L. Looger. Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators. *Nat. meth.*, 6:875–81+, 2009.

L. Tian, S.A. Hires, and L.L. Looger. Imaging neuronal activity with genetically encoded calcium indicators. *Cold Spring Harb Protoc.*, 6:647–656, 2012.

I. Tsamardinos, C. F. Aliferis, and A. Statnikov. Time and sample efficient discovery of markov blankets and direct causal relations. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 673–678, 2003.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Machine Learning*, 65(1):31–78, 2006.

Marx V. Cancer genomes: discerning drivers from passengers. *Nat. Meth.*, 11(4): 375–379, 2014.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI'90, pages 255–270, 1991.

B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler. Cancer genome landscapes. *science*, 339(6127):1546–1558, 2013.

J. Watkinson, K.-C. Liang, X. Wang, T. Zheng, and D. Anastassiou. Inference of Regulatory Gene Interactions from Expression Data Using Three-Way Mutual Information. *Annals of the New York Academy of Sciences*, 1158:302–313, March 2009.

N. K. Wilson, R. T. Timms, S. J. Kinston, Y. H. Cheng, S. H. Oram, J. R. Landry, J. Mullender, K. Ottersbach, and B. Gottgens. Gfi1 expression is controlled by

five distinct regulatory regions spread over 100 kilobases, with scl/tal1, gata2, pu.1, erg, meis1, and runx1 acting as upstream regulators in early hematopoietic cells. *Mol. Cell. Biol.*, 30(15):3853 – 3863, 08 2010.

C. Wu, C. Orozco, J. Boyer, M. Leglise, J. Goodale, S. Batalov, C.L. Hodge, J. Haase, J. Janes, 3rd J. W. Huss, et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol*, 10 (11):R130, 2009.

R. Yuste and L. C. Katz. Control of postsynaptic ca2+ influx in developing neocortex by excitatory and inhibitory neurotransmitters. *Neuron*, 6:333–344, 1991.

J. Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.*, 172(16-17): 1873–1896, 2008a.

J. Zhang. Causal reasoning with ancestral graphs. *J. Mach. Learn. Res.*, 9:1437–1474, June 2008b.

W. Zhao, E. Serpedin, and E. R. Dougherty. Inferring connectivity of genetic regulatory networks using information-theoretic criteria. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(2):262–274, 2008.

M. K. Zia, K. A. Rmali, G. Watkins, R. E. Mansel, and W. G. Jiang. The expression of the von hippel-lindau gene product and its impact on invasiveness of human breast cancer cells. *Int. J. Cancer.*, 20:605–11, 2007.

F. Zohren, G. P. Souroullas, M. Luo, U. Gerdemann, M. R. Imperato, N. K. Wilson, B. Göttgens, G. L. Lukov, and M.A. Goodell. The transcription factor lyl-1 regulates lymphoid specification and the maintenance of early t lineage progenitors. *Nat. Immunol.*, 13(8):761 – 769, 2012.