Annotation du proteome d'A. thaliana via l'analyse et la prédiction de son interactome

Simon Gosset

université PARIS-SACLAY

Annotation du proteome d'*A. thaliana* via l'analyse et la prédiction de son interactome

Annotation of the A. thaliana proteome via the analysis and prediction of its interactome

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 577, Structure et Dynamique des Systèmes Vivants (SDSV)
Spécialité de doctorat : Biologie Computationnelle
Graduate school : Life Sciences and Health
Référent : Université d'Évry Val d'Essonne

Thèse préparée dans l'unité de recherche **Institute of Plant Sciences Paris-Saclay (IPS2) (Université Paris-Saclay, CNRS, INRAE, Univ Evry)** sous la direction de **Marie-Hélène MUCCHIELLI-GIORGI**, professeure

Thèse soutenue à Paris-Saclay, le 9 février 2024, par

Simon GOSSET

Composition du jury

Membres du jury avec voix délibérative

Blaise HANCZAR Professeur, Université d'Evry	Président
Juliette MARTIN Directrice de Recherche, CNRS, ENS de Lyon	Rapportrice
Thomas SCHIEX Directeur de Recherche, INRAE, MIA Toulouse	Rapporteur
Jessica ANDREANI Chargée de recherche, CEA, I2BC	Examinatrice
Martin WEIGT Professeur, Sorbonne Université	Examinateur

THESE DE DOCTORAT
NNT : 2024UPASL004

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Annotation du proteome d'*A. thaliana* via l'analyse et la prédiction de son interactome

Mots clés : Interaction protéine-protéine, annotation, prédiction, réseau, graphe, AlphaFold

Résumé : Le fonctionnement des cellules vivantes est assuré par un ensemble d'interactions entre des molécules que l'on appelle protéines. Identifier les couples de protéines en interaction, impliqués dans un processus biologique d'intérêt, permet donc de mieux comprendre son fonctionnement. Pour cela, il existe un ensemble de méthodes expérimentales, mais qui sont trop coûteuses pour explorer l'ensemble des interactions mises en jeu ou pas passez fiables. Pour pallier à ce problème, des méthodes informatiques ont été développées. L'objectif de ma thèse a ainsi été de mettre en place une méthode permettant de construire un réseau d'interactions protéine-protéine (PPI) impliquant des protéines d'intérêt, puis d'y rechercher toutes les protéines impliquées dans un même processus biologique qui constituent des sous-réseaux de protéines fortement interconnectées. Le résultat dépendant de la qualité du réseau, j'ai tenté ensuite de l'affiner en y ajoutant des PPI prédites. Pour cela, j'ai mis en place une méthode de prédiction de PPI utilisant Alphafold-multimer, une méthode innovante de prédiction de structure de complexe protéique. En parallèle, j'ai produit un jeu de données décrivant les caractéristiques physico-chimiques, l'énergie de liaison et la propension à l'interaction des surfaces et des interfaces d'un grand nombre de protéines afin de comprendre ce qui distingue les PPI correctement prédites de celles ratées par la méthode que j'ai mise en place.

Title : Annotation of the *A. thaliana* proteome via the analysis and prediction of its interactome

Keywords : Protein-protein interaction, annotation, prediction, network, graph, AlphaFold

Abstract : Living organisms function thanks to a set of interactions between molecules called proteins. Identifying the pairs of proteins involved in these interactions provides a better understanding of how they function. There are several experimental methods for identifying these proteins, but they are either too expensive or not reliable enough to explore all the protein interactions that take place in living organisms. Computational methods have been developed to predict these interactions to meet this need. During my thesis, I set up a method for predicting protein-protein interactions using Alphafold-multimer, an innovative structure prediction method. At the same time, I produced a dataset describing the physico-chemical characteristics, binding energy and interaction propensity of the surfaces and interfaces of a large number of proteins to understand what distinguishes correctly predicted PPIs from incorrectly predicted PPI by the method I have set up.

Table des matières

Remerciements

J'aimerai commencer par remercier Marie-Hélène, qui m'a fait confiance depuis le début de mon stage de master 2 jusqu'à la fin de ma thèse, et sur qui j'ai toujours pu compter. Je la remercie pour nos échanges sincères et pour m'avoir soutenu dans les choix que j'ai fait au cours de la thèse. A son contact, j'ai beaucoup appris aussi bien scientifiquement que humainement.

Je remercie aussi tout les membres de l'équipe Gnet pour les moments conviviaux passés au café et pour les échanges scientifiques fructueux lors des réunions d'équipe. Je remercie en particulier Marie-Laure qui a toujours su se rendre disponible pour discuter avec moi lorsque j'ai eu des doutes et qui m'a soutenu dans les moment difficiles. Je remercie aussi Jean-Philippe qui m'a notamment aidé à lancer des calculs sur la fin de ma thèse.

J'aimerai remercier Anne Lopes de l'I2BC et Hugo Schweke pour l'aide qu'ils m'ont apportée sur le projet IPOPS.

Je tiens à exprimer ma gratitude envers Annie Glatigny pour le soutien précieux qu'elle m'a prodigué dans le cadre du projet APPINetwork.

Je remercie Anthony et Amiel pour le travail qu'ils ont réalisé au court de leur stage.

Je remercie le service Bio-informatique de l'IPS2, et en particulier Frédéric pour l'aide technique qu'il m'a apportée à de nombreuses reprises, et pour sa rapidité à résoudre mes problèmes.

Je remercie également les membres de mon jury de thèse pour avoir accepté d'évaluer mon travail. Je remercie Thomas Schiex et Juliette Martin d'avoir accepté d'être rapporteurs de ma thèse. Je les remercies pour leur remarques constructives et idées très pertinentes qui m'ont permis d'améliorer mon manuscrit . Je remercie Jessica Andreani, Martin Weigt et Blaise Hanczar pour avoir accepté respectivement d'examiner et de présider ma thèse. Je remercie aussi les membres de mon comité de thèse, Raphael Guerrois, Isabelle BLOCH et Michel Zivy d'avoir suivi l'avancée de mes travaux pendant ces trois ans et de m'avoir aidé à faire des choix.

Enfin, pour terminer, je remercie mes parents, mon frère et ma soeur ainsi que l'ensemble de ma famille pour leur amour et leur soutien, ainsi que tous mes amis qui m'ont donné la force et le courage d'affronter cette épreuve qu'est la thèse.

1 - Les protéine molécules fonctionnelles du vivant

1.1 . Qu'est-ce qu'une protéine ?

1.1.1 . Introduction/historique

Les protéines sont des éléments essentiels à la vie, jouant un rôle fondamental dans de nombreux processus biologiques nécessaires au fonctionnement des organismes vivants. Elles sont composées d'une séquence d'acides aminés qui se lient les uns aux autres par une liaison appelée liaison peptidique. Il existe 20 acides aminés standard qui constituent la majorité des protéines présentes dans les organismes vivants. Ces acides aminés s'organisent dans l'espace pour former des structures tridimensionnelles. Ces structures peuvent ensuite s'assembler pour former des complexes ou interagir avec leurs cibles afin d'accomplir les fonctions nécessaires au bon fonctionnement des cellules. La compréhension du fonctionnement des protéines, ainsi que de leurs interactions, a été et constitue encore un objectif central de la recherche en biologie.

Au XVIII^{ème} siècle, les protéines ont été pour la première fois reconnues comme une classe distincte de molécules biologiques, grâce notamment aux travaux d'Antoine Fourcroy. Cependant, ce n'est qu'en 1838 que le terme "protéine" a été utilisé pour la première fois dans un article, grâce à Gerhardus Johannes Mulder ([43]), suite à la suggestion du chimiste Jöns Jacob Berzelius. Dans cet article, Mulder a observé que toutes les protéines avaient en commun une partie de leur composition. Convaincu que les protéines étaient une substance fondamentale pour la vie, le terme "protéine" a été choisi en référence au mot grec "protéios", signifiant "premier" ou "primordial".

En 1894, Emil Fischer, propose une première théorie concernant l'interaction des protéines avec leur cible : le modèle clé-serrure. Ce modèle propose que les structures tridimensionelles d'une protéine et de sa cible doivent être complémentaires pour qu'elles puissent interagir l'une avec l'autre, tel la forme d'une clé rentrant dans une serrure [69].

La découverte des 20 acides aminés standards constituant les protéines s'est étalée sur plus d'un siècle. En 1806, l'asparagine fut le premier acide aminé isolé à partir d'extraits d'asperge par le chimiste français Nicolas-Louis Vauquelin [73]. La thréonine est le dernier acide aminé standard à être découvert en en 1935 par William Cumming Rose [38]. C'est Emil Fischer qui avança l'idée que les acides aminés étaient reliés entre eux par une liaison amine, qu'il nomma la liaison peptidique, for-

mant ainsi une séquence d'acides aminés. En 1907, il parvint à prouver cette théorie en synthétisant une séquence de 18 acides aminés et en démontrant que celle-ci pouvait être clivée par des enzymes de la même manière que les protéines naturelles. Il faudra cependant attendre 1949 avant que pour la première fois, la séquence d'acides aminés d'une protéine soit entièrement déterminée par Fréderick Sanger [57].

En 1958, Daniel Koshland propose une nouvelle théorie concernant la manière dont les protéines interagissent entres elles : l'"induced-fit". Cette théorie s'opposant au modèle clé-serrure est aujourd'hui largement acceptée et propose que les structures des protéines ne soient pas entièrement rigides. Ainsi lorsque 2 protéines interagissent l'une avec l'autre, leurs structures tridimensionnelles se modifient légèrement afin de stabiliser l'interaction [69].

La même année, une percée majeure a lieu dans le domaine. La première structure tridimensionnel de protéine est résolue par John Kendrew et Max Perutz par cristallographie aux rayons X, c'est la structure de la myoglobine du grand cachalot, une protéine présente dans les muscles [25]. C'est le début de l'étude de la structure des protéines. En 1971, une base de donnée est fondée pour accueillir les structures protéiques déterminées expérimentalement : la Protein Data Bank (PDB). Au début elle ne contient que 5 structures, mais ce nombre va rapidement évoluer. En 1981, elle en contenait déjà 100, puis en 1991, elle en contenait 1 000. En 2001, elle en contenait 10 000 [3]. Et en 2021, elle en contenait plus de 170 000 [2]. L'évolution rapide du nombre de structures présentes dans la PDB est due à l'évolution des techniques de résolution des structures, notamment la RMN et la microscopie électronique, mais aussi au développement des moyens informatiques.

Malgré toutes ces évolutions, les résolutions expérimentales restent longues et coûteuses. C'est pourquoi de nombreux efforts ont été déployés dans le développement de méthodes *in silico*. En 2021, une méthode de prédiction des structures est publiée et révolutionne le domaine de la biochimie structurale, Alphafold2 [22]. Elle permet la prédiction de structure d'une qualité très proche des méthodes expérimentales. Une base de données est mise à disposition de la communauté scientifique et contient plus de 200 millions de structures (en juillet 2022). Cette nouvelle méthode a révolutionné le domaine et offre de nouvelles opportunités qu'on ne pensait pas possible il y a quelques années, notamment pour la prédiction et l'étude des interactions entre les protéines.

1.1.2 . La structure des protéines

La structure des protéines conditionne leurs propriétés et leur fonction. On distingue quatre niveaux d'organisation structurale. Chaque

niveau est essentiel pour assurer le bon fonctionnement des protéines dans l'organisme. Des modifications au niveau de la structure peuvent entraîner des dysfonctionnements cellulaires.

La structure primaire des protéines est le premier niveau d'organisation. Elle représente la séquence linéaire des acides aminés qui composent une protéine. Cette séquence est déterminée par le code génétique contenu dans l'ADN, où chaque triplet de nucléotides code pour un acide aminé particulier.

Les acides aminés sont les briques de base qui composent les protéines. Il existe 20 acides aminés standards. Les acides aminés possèdent une partie commune constituée d'un groupe amine (NH_2), un groupe carboxyle (COOH), une partie qu'on appelle le groupe « radical » ou « R » qui est spécifique à chaque acide aminé et lui confère des propriétés physicochimiques particulières et d'un carbone alpha central lié à un hydrogène et au groupe amine, carboxyl et R.

La diversité des acides aminés et leur ordre séquentiel unique confèrent à chaque protéine une identité et une fonction distincte dans l'organisme. Des liaisons covalentes, appelées liaisons peptidiques, connectent les acides aminés successifs par une réaction entre le groupe carboxyl d'un acide aminé et le groupe amine d'un autre acide aminé avec élimination d'une molécule d'eau. Cela forme ainsi une longue chaîne qui constitue le "squelette" de la protéine que l'on appelle la chaîne principale. La structure primaire joue un rôle fondamental dans la détermination des autres niveaux de structure.

La structure secondaire est la configuration locale de la chaîne protéique. Elle représente un premier niveau d'organisation de la chaîne d'acides aminés dans l'espace. Le plus souvent, on observe des structures secondaires en hélices alpha ou en feuillets bêta.

L'hélice alpha est une structure hélicoïdale où la chaîne principale s'enroule autour d'un axe central. Cette structure se stabilise grâce à des liaisons hydrogène formées entre les atomes d'oxygène engagés dans les liaisons peptidiques et les atomes d'hydrogène du groupe amine d'un acide aminé voisin, situé quatre positions plus loin dans la séquence. Cette disposition permet une stabilité accrue et confère à l'hélice alpha une certaine rigidité structurelle. On la retrouve fréquemment dans les fragments transmembranaires des protéines membranaires.

Le feuillet beta est une structure qui se forme lorsque la chaîne principale se replie en plusieurs brins. Ces brins peuvent être antiparallèles (es brins s'étendent dans des directions opposées), ou parallèles (es brins s'étendent dans la même direction). Cette structure est stabilisée par

des liaisons hydrogène formées entre les groupes amines et les groupes carbonyles des liaisons peptidiques (CO) des brins adjacents, donnant lieu à une structure en feuillet qui est très résistante.

La structure tertiaire représente le repliement complet d'une protéine en 3 dimensions. Les structures secondaires s'organisent pour former des domaines fonctionnels, c'est à dire des régions stables avec des fonctions distinctes. Ces domaines sont généralement séparés par des boucles plus flexibles permettant l'arrangement globale des domaines. La stabilisation de la structure tertiaire fait intervenir des liaison de différentes natures entre les atomes de la chaîne latérale des acides-aminés : les liaisons hydrogène, les interactions électrostatiques, les interactions hydrophobes et les ponts disulphures. Ce repliement tridimensionnel est très fortement lié à la fonction de la protéine. Un repliement incorrect de la protéine peut ainsi mener à une protéine défectueuse et peut être responsable de maladie grave (par exemple les maladies à prions).

Le dernier niveau d'organisation des protéines est la structure quaternaire, il représente l'organisation de plusieurs protéines que l'on va alors appeler les sous-unités, qui vont s'associer les unes avec les autres afin de former des complexes protéiques fonctionnels. Ces sous-unités peuvent être identiques (homomères) ou différentes (hétéromères) et s'assemblent grâce à des liaisons non covalentes, telles que les liaisons ioniques, hydrogène et hydrophobes.

Des exemples célèbres de structures quaternaires sont :

- l'hémoglobine, une protéine présente dans les globules rouges, qui est un tétramère (4 sous unités protéiques) composé de deux sous-unités alpha et deux sous-unités bêta.

- l'ADN polymérase, une enzyme impliquée dans la réplication de l'ADN, un processus biologique essentiel à la vie, et qui est composée de multiples sous-unités.

La compréhension des structures quaternaires, c'est-à-dire des complexes formés par des protéines monomériques est essentielle pour élucider les mécanismes moléculaires complexes impliqués dans les processus biologiques.

1.2 . Les interactions protéine-protéine au cœur des proces-

sus biologiques

La notion d'interaction protéine-protéine (PPI), c'est à dire l'interaction physique entre protéines, est au coeur du fonctionnement des processus biologiques. Comprendre comment les protéines interagissent entre elles permet de mieux appréhender le fonctionnement du vivant et représente aujourd'hui un domaine de recherche très actif.

Les interactions protéine-protéine jouent un rôle crucial dans divers processus biologiques, notamment en modulant l'activité des enzymes. Un exemple significatif de cette régulation se produit lors du contrôle du cycle cellulaire, un processus essentiel au fonctionnement et au développement de tout organisme vivant. La progression entre les différentes étapes de ce cycle doit être soigneusement orchestrée. Les protéines CDK (kinase dependent protein) jouent un rôle essentiel dans ce contrôle. Leur activité kinase est initialement très faible lorsqu'elles sont isolées. Cependant, lorsqu'elles interagissent avec une autre famille de protéine, les cyclines, leur activité kinase est activée [34].

Les PPI jouent un rôle crucial dans la transduction du signal, par exemple chez les plantes, qui, étant des organismes immobiles, doivent répondre rapidement et efficacement aux stress environnementaux pour assurer leur survie. Dans ce contexte, les protéines kinases activées par les mitogènes (MAPK) jouent un rôle clé dans la propagation du signal lorsqu'un stimulus extracellulaire est détecté. La transmission et l'amplification du signal sont assurés par des cascades de phosphorylation, c'est-à-dire que des séries de sous-familles des MAPK sont activées les unes après les autres notamment par l'ajout d'un groupe phosphate. Ainsi si une MAPKKK active 3 MAPKK, puis que chacune de ces MAPKK active 3 MAPK... il ya bien une transmission et une amplification de ce signal [66]. Pour que ces phosphorylations se produisent, il est essentiel que les protéines impliquées puissent interagir les unes avec les autres. Identifier les cibles en aval des MAPK dans la cascade, c'est-à-dire les protéines avec lesquelles les MAPK interagissent, permettrait ainsi de mieux comprendre les mécanismes de réponse au stress chez les plantes. En identifiant ces interactions protéine-protéine, nous pourrions découvrir les régulations fines qui contrôlent les réponses cellulaires au stress environnemental. Une meilleure compréhension de ces mécanismes est d'une importance capitale pour améliorer la résistance des plantes aux stress environnementaux. Cela contribue ainsi au développement d'une agriculture qui assurerait la sécurité alimentaire face aux défis du changement climatique.

Les PPI sont essentielles à la formation de complexes protéiques fonctionnels. Comme mentionné lors de la description de la structure quaternaire, certains complexes protéiques mettent en jeu des interactions entre plusieurs sous-unités protéiques. Ces sous-unités doivent interagir

avec les bons partenaires, aux bons endroits, afin que le complexe final soit fonctionnel. On peut citer comme exemple l'ARN polymérase II, une enzyme responsable de la transcription des gènes en ARN messager. Le cœur de ce complexe serait composé de 12 sous-unités protéiques chez la levure *S. cerevisiae*. Des PPI supplémentaires ont été identifiés et sont nécessaires pour que l'enzyme puisse reconnaître spécifiquement les promoteurs des gènes à exprimer [44].

Les PPI participent aussi au maintien de l'intégrité structurelle des cellules. Un exemple notable est celui des protéines du cytosquelette, en particulier les filaments intermédiaires. Présentes chez la majorité des vertébrés et des invertébrés, ces protéines interagissent les unes avec les autres, s'organisant en fibres d'un diamètre d'environ 10 nm. Grâce à leur stabilité et leur résistance, ces filaments intermédiaires contribuent activement à la cohésion des tissus et à la résistance mécanique des cellules [17].

Enfin, il est très important de mentionner que les protéines évoluent dans un environnement très dense en protéines. Les interactions physiques entre protéines ne se limitent donc pas à des interactions participant activement à une fonction biologique. De plus, une fraction non négligeable de paires de protéines (estimé à 8.7%) aurait une structure compatible pour former une interaction stable. Cela rend l'identification des PPI fonctionnelles difficile, d'autant plus que la fraction de PPI fonctionnelles relativement au nombre de couples de protéines qu'il est possible de former est faible (estimée aux alentours de 0.2%) [30].

1.2.1 . Les caractéristiques des PPI

Les PPI peuvent être caractérisées en fonction de :

- leur stabilité : on va pouvoir distinguer les interactions permanentes et les interactions transitoires. Les interactions permanentes se produisent lorsque les partenaires protéiques s'associent de manière durable pour remplir leur fonction. C'est le cas des protéines qui s'associent pour former des complexes fonctionnels.

  En revanche, les interactions transitoires sont des liaisons plus temporaires et dynamiques entre protéines. Elles peuvent être plus faibles que les interactions permanentes et sont souvent régulées de manière précise dans le temps et l'espace. Les interactions transitoires jouent un rôle clé dans de nombreux processus cellulaires tels que la signalisation cellulaire, la régulation de l'expression génique et le transport intracellulaire. Ces interactions permettent aux protéines de s'assembler et de collaborer, en réponse aux signaux environnementaux ou aux étapes du cycle cellulaire. La coexistence de ces deux types d'interaction, permanente et

transitoire, confère une grande flexibilité aux réseaux protéiques et permet leur régulation. Les interactions permanentes fournissent la stabilité et la cohésion nécessaires à la formation de complexes protéiques fonctionnels, tandis que les interactions transitoires permettent des réponses rapides et réversibles aux signaux cellulaires. Enfin il est important de rajouter que les portéines ne tombent pas directement dans une catégorie ou dans l'autre mais qu'il existe plutôt un gradient continue entre ces deux catégories.

- la spécificité entre les deux partenaires en interaction : Les protéine évoluent dans un environnement encombré avec de nombreux partenaires potentiels. Pour remplir leur fonction biologique de manière efficace, de nombreuses protéines présentent une grande spécificité dans leur choix de partenaires, leur permettant d'interagir avec le bon partenaire au sein de l'environnement dense de la cellule. Cependant, certaines protéines sont multispécifiques, ce qui signifie qu'elles peuvent interagir avec des protéines appartenant à différentes familles de protéines. Ce phénomène est courant dans les processus de signalisation intra et extracellulaire. Cette spécificité en terme de partenaire dépend de la complémentarité de structure et des caractéristiques physico-chimique des partenaires, mais aussi de la localisation cellulaire des protéines, et du moment où celles-ci sont exprimées [21], puisque pour que deux partenaires puissent interagir, ceux-ci doivent être exprimés au même moment au même endroit de la cellule.

- si l'interaction est obligatoire ou non obligatoire : Une PPI est dite obligatoire lorsque les deux protéines impliquées ne peuvent pas être trouvées séparément dans une cellule vivante. Ces interactions sont souvent permanentes. Une PPI est dite non obligatoire lorsque les deux protéines impliquées peuvent être trouvées séparément dans une cellule vivante [21].

- la nature des partenaires qui participe à l'interaction : On peut distinguer les homo-oligomères, qui sont des complexes composés de sous-unités protéiques identiques, et les hétéro-oligomères, qui sont des complexes composés de sous-unités protéiques différentes [21].

- si les partenaires sont en contact ou non : Une interaction directe est une interaction dans laquelle les deux protéines impliquées sont en contact physique. Une interaction indirecte est une inter-

action dans laquelle les deux protéines impliquées ne sont pas en contact physique, mais sont néanmoins présentes au sein du même complexe [37]. Il est important de distinguer les interactions directes et indirectes car certaines méthodes qui identifient des PPI ne peuvent pas déterminer si l'interaction est directe ou indirecte. C'est par exemple le cas de la co-immunoprécipitation (co-IP) [32].

1.2.2 . Les forces qui régissent les PPI

La diversité de caractéristiques dont peuvent faire preuve les PPI peut être expliquée par les différentes forces qui les régissent.

1.2.2.1 . La force électrostatique

La force électrostatique est la résultante des charges présentes sur les atomes. Lorsqu'un atome est chargé négativement et un autre positivement, ils sont mutuellement attirés, tandis que des charges similaires se repoussent. Les acides aminés constitutifs des protéines diffèrent par les propriétés de leurs chaînes latérales. Certaines chaînes latérales d'acides aminés sont chargées négativement, tels que l'acide aspartique et l'acide glutamique, en raison de leurs groupements acide-carboxylique. D'autres présentent une charge positive, comme l'arginine, la lysine et l'histidine, liées à la présence d'atomes d'azote sur leurs chaînes latérales. Lorsque deux acides aminés sont distants d'environ 10 Å et diffèrent par une seule charge nette, la force électrostatique devient prépondérante dans leur interaction. Cependant, la force électrostatique est également impliquée dans des interactions à plus longue portée, contribuant à la reconnaissance et au guidage mutuel des partenaires [71].

1.2.2.2 . Les liaisons hydrogène

Les liaisons hydrogène, en grande partie d'origine électrostatique [36], englobent l'interaction entre un atome d'hydrogène lié à un atome électronégatif, généralement de l'oxygène, et un autre atome électronégatif. Considérons la molécule d'eau H_2O comme exemple. L'atome d'oxygène, en liaison avec les deux atomes d'hydrogène, est plus électronégatif, ce qui rapproche les électrons des liaisons covalentes de l'oxygène. Cette polarisation confère une charge partielle positive aux hydrogènes et deux charges partielles négatives à l'oxygène. Les atomes d'hydrogène d'une molécule d'eau peuvent former des liaisons hydrogène avec les atomes d'oxygène d'autres molécules d'eau. Cette dynamique est également présente dans les acides aminés, impliquant les groupes dans les liaisons peptidiques entre le groupe N-H de la chaîne principale d'un acide aminé et le groupe C=O d'un autre, favorisant la stabilité des structures secondaires [45]. Elle se retrouve aussi au niveau des chaînes laté-

rales d'acides aminés, qui peuvent interagir avec des molécules d'eau ou d'autres chaînes latérales, contribuant aux interactions protéine-protéine et à la consolidation des complexes. Bien que chaque liaison hydrogène individuelle ne soit pas très robuste, leur accumulation au sein des complexes protéiques assure leur stabilité et leur forme [19]. Les acides aminés pouvant former des liaison hydrogène via leur chaine latérale sont les acides aminés polaires, acide aspartique et acide glutamique, ainsi que les acides aminés polaires non chargés à savoir la sérine, la thréonine, la cystéine, la tyrosine, l'asparagine et la glutamine.

1.2.2.3 . La force hydrophobe

Contrairement à la force électrostatique, la force hydrophobe est causée par l'absence de charge de certaines chaînes latérales des acides aminés. En effet, les régions hydrophobes de certaines protéines ont une tendance naturelle à s'associer avec d'autres régions hydrophobes, car cela réduit leur exposition à l'eau environnante. Les liaisons hydrogène formées par les molécules d'eau sont bien plus favorables que les interactions possibles avec des atomes non chargées. Ainsi lorsque deux protéines interagissent l'une avec l'autre, les régions hydrophobes vont avoir tendance à s'orienter l'une vers l'autre pour réduire leur exposition à l'eau, jouant ainsi un rôle important dans la stabilisation des complexes protéiques. Les acides aminés hydrophobes sont la glycine, l'alanine, la valine, l'isoleucine, la leucine, la méthionine, la phénylalanine et la proline, liée au fait que leur chaîne latérale ne peut pas faire de liaison hydrogène et ne sont pas chargée [7].

1.2.2.4 . La force de Van der Waals (vdw)

Les forces de Van der Waals résultent de fluctuations naturelles des charges électroniques, créant des variations temporaires dans la distribution des charges électriques au sein des molécules. En conséquence, des régions momentanément positives et négatives apparaissent, générant une interaction attirante entre les régions positives d'une molécule et les régions négatives d'une autre. Les forces de van der Waals sont particulièrement pertinentes dans les interactions entre les chaînes latérales des acides aminés, où les groupes apolaires peuvent entrer en contact étroit. Ces interactions peuvent être cruciales pour l'ajustement précis des surfaces des protéines, favorisant l'emboîtement des structures protéiques et contribuant ainsi à la formation de complexes fonctionnels. Bien que les forces de van der Waals soient généralement considérées comme relativement faibles, leur accumulation au sein de l'interface entre deux protéines peut aboutir à des interactions globalement stables et spécifiques [56].

1.2.3 . Identifier les PPI expérimentalement

L'identification de l'interactome (l'ensemble des PPI) d'une espèce est un problème difficile à cause de sa grande dimensionalité. En effet, prenons l'exemple de la plante modèle *Arabidopsis thaliana* : elle compte 27 481 gènes codant pour 39 319 protéines dans son protéome de référence Uniprot. Si l'on cherche à identifier l'ensemble de l'interactome de cette plante sans *a priori* et par paire, ce ne sont pas moins de 772 972 221 paires de protéines qu'il faut tester pour vérifier les interactions. D'autre part, la taille de l'interactome de *A. thaliana* est estimée à un peu plus de 200 000 PPI [33]. Par conséquent, la proportion de PPI par rapport au nombre de paires possibles est très faible.

De nombreuses méthodes ont été développées pour identifier les PPI. Ces approches peuvent être classées en deux catégories distinctes : les méthodes à bas débits (résonance magnétique nucléaire, microscopie électronique, cristallographie aux rayons X...), coûteuses et difficiles à mettre en place, et les méthodes à haut débit, permettant de générer un grand volume de données mais de moins bonne qualité.

1.2.3.1 . Les méthodes à bas débit

Les approches à bas-débit, bien que souvent associées à des coûts et une complexité accrue pour leur mise en œuvre, permettent d'identifier les PPI avec une grande fiabilité. Ces méthodes ciblent généralement les PPI pour lesquelles des indices préalables suggèrent une interaction entre les protéines impliquées.

La cristallographie aux rayons X est une méthode fondamentale dans l'étude de la structure des complexes protéiques. Elle permet d'obtenir des informations détaillées sur la structure tridimensionnelle de protéines d'intérêt, à condition de réussir à obtenir la protéine ou le complexe d'intérêt sous forme de cristal. Suite à cela, la protéine est exposée à des rayons X, qui sont diffractés par les atomes présents dans le cristal. L'analyse des schémas de diffraction résultants permet de déterminer la disposition spatiale des atomes dans la protéine, révélant ainsi sa structure. Utilisée pour identifier la structure des complexes protéiques, cette méthode fournit des informations précieuses sur les sites d'interaction, les conformations des protéines en interaction et les détails moléculaires des liaisons entre les partenaires protéiques [6]. Cependant, l'obtention du cristal est généralement empirique et difficile, notamment lorsque les interactions protéine-protéine sont faibles, transitoires ou flexibles. De plus, la technique ne permet pas de capturer les états dynamiques et les variations conformationnelles des protéines en interaction [9, 41] ce qui n'est pas forcément le cas de méthodes plus récentes tel que la cryo-microscopie électronique.

En effet, la cryo-microscopie électronique est une technique expérimentale puissante utilisée pour étudier les complexes biologiques à l'échelle moléculaire. Elle permet de visualiser en haute résolution des structures tridimensionnelles de protéines seules ou en complexes, tout en préservant leur état naturel. Dans cette méthode, les échantillons biologiques sont congelés rapidement à des températures très basses pour figer leur structure dans leur conformation biologique. Cette étape évite l'utilisation de la fixation chimique classiquement employée pour immobiliser et conserver les échantillons biologiques, ce qui pourrait altérer la structure native des protéines. Ensuite, des images 2D de l'échantillon sont prises sous différents angles à l'aide d'un microscope qui bombarde les échantillons avec des électrons. Ces images sont ensuite combinées informatiquement pour reconstruire une image 3D détaillée du complexe étudié. La cryo-microscopie électronique est particulièrement utile pour l'étude de complexes de grande taille et flexibles. En capturant la structure des complexes dans différentes conformations, elle a permis des avancées significatives dans la compréhension des mécanismes moléculaires et des interactions au sein des cellules et des organismes [6, 82]. Néanmoins la résolution de la cryo-miscroscopie électronique n'est pas toujours aussi bonne que celle de la cristallographie aux rayons X.

1.2.3.2 . Les méthodes à haut débit

Les approches à haut débit permettent d'explorer un grand nombre de PPI potentielles et donc d'explorer une partie bien plus importante d'un interactome (c'est à dire l'ensemble des interactions entre les protéines d'un organisme) que les méthodes bas débit, néanmoins, les résultats sont moins fiables.

La méthode du double hybride représente une approche à haut débit largement employée pour identifier les interactions protéine-protéine (PPI) entre des paires d'interactants. Cette approche expérimentale est réalisée *in vivo* chez la levure *S. cerevisiae* et repose sur la transcription d'un gène rapporteur, un gène dont la transcription peut être facilement observée. Afin d'activer la transcription du gène rapporteur, cela nécessite la proximité d'un domaine de liaison à l'ADN (BD) et d'un domaine d'activation (AD). Pour déterminer la présence d'une PPI entre deux protéines d'intérêt, le domaine BD est fusionné à l'une des protéines, appelée "bait", c'est à dire que l'on ajoute la séquence en acides aminés du BD au début où à la fin de la protéine bait, tandis que le domaine AD est fusionné à l'autre protéine, appelée "prey". Si le bait et la prey interagissent physiquement, l'AD et le BD se rapprochent suffisamment pour induire l'expression du gène rapporteur. Bien que le double hybride soit en mesure d'évaluer un vaste nombre de paires de protéines *in vivo*,

il n'est pas exempt de limitations. Tout d'abord, cette méthode est menée au sein de la levure *S. cerevisiae*, même lorsque les protéines testées proviennent d'organismes différents. Les compartiments cellulaires dans lesquels les interactions sont testées ne correspondent pas nécessairement aux compartiments où les protéines sont exprimées dans des conditions biologiques normales. De plus, cette méthode ne prend pas en compte les régulations qui pourraient intervenir dans des conditions biologiques réelles. En conséquence, cette approche peut engendrer des faux positifs, car elle pourrait identifier une interaction entre des protéines qui ne se rencontrent jamais dans des conditions biologiques réelles, soit parce qu'elles ne sont pas exprimées simultanément en raison de régulations, soit parce qu'elles ne sont pas présentes dans les mêmes compartiments cellulaires.

Une autre méthode à haut débit est le tandem Affinity Purification-Mass Spectrometry (TAP-MS). Cette méthode permet de purifier des complexes protéiques dans leur état natif, puis d'analyser ces complexes grâce à la spectrométrie de masse. Initialement, une protéine d'intérêt est choisie et marquée avec deux étiquettes distinctes par fusion génétique. Cette protéine marquée est appelée "bait". Les cellules produisant cette protéine fusionnée sont ensuite lysées, et les extraits cellulaires obtenus sont soumis à une colonne de chromatographie qui retient spécifiquement la première étiquette de la protéine "bait". De cette manière, les composants cellulaires qui ne sont pas en interaction avec la protéine "bait" ne sont pas retenus par la colonne. La protéine "bait" et les molécules associées sont ensuite isolées séparément et soumises à une deuxième étape de purification à travers une colonne de chromatographie d'affinité, capable de retenir la deuxième étiquette. Après cette étape, la protéine "bait" et tout ce qui lui est lié sont récupérés pour être ensuite analysés par spectrométrie de masse. La spectrométrie de masse est une méthode largement utilisée en protéomique, qui permet d'analyser la masse et la composition des molécules présentes dans un échantillon. En utilisant la spectrométrie de masse, les partenaires protéiques qui sont associés à la protéine "bait" peuvent être identifiés. Les deux passages dans des colonnes de chromatographie permette d'obtenir des interactions avec une spécificité élevée. De plus cette méthode permet d'obtenir la composition des complexes biologiques auxquelles la protéine bait participe. Néanmoins, une limitation de cette méthode réside dans le fait qu'elle ne permet pas de distinguer clairement si les interactions identifiées sont directes ou indirectes. Il peut être difficile de déterminer si les composants des complexes interagissent directement ou à travers d'autres protéines intermédiaires. De plus, cette méthode est moins ap-

propriée pour mettre en évidence des interactions qui sont temporaires ou transitoires.

1.2.3.3 . Les difficultés rencontrées par l'identification expérimentale

Les difficultés rencontrées par l'identification expérimentale des PPI résident en premier lieu dans la nature chronophage, exigeante en expertise et coûteuse de ces méthodes. D'autant plus qu'il y a un vaste volume de données à examiner. Ces approches sont également assez imprécise pour la détection de PPI transitoires. Cette imprécision peut conduire à un biais dans les interactions mises en évidence, en privilégiant les interactions stables plus faciles à détecter [10]. Néanmoins, la plus grosse difficulté rencontrée actuellement réside sans doute dans les méthodes à haut débit, les méthodes générant le plus de données. Ces méthodes ont tendance à produire des faux-positifs et des faux-négatifs, menant à de nombreuses incohérences sur les résultats qui sont produits. Ainsi en fonction de la méthode utilisée pour identifier une PPI, les résultats obtenu pour la même paire de protéine peuvent être différents [13, 18, 62].

1.2.4 . Explorer les interactomes in silico

L'interactome est un terme qui désigne l'ensemble des interactions entre les protéines d'un organisme. L'étude des interactomes *in silico* passe généralement par la construction d'un réseau d'interactions protéine-protéine. La construction de ce réseau nécessite de pouvoir récupérer les données de PPI expérimentales dans des bases de données de PPI. Au vu des problèmes rencontrés par l'identification des PPI expérimentalement, la construction d'un interactome complet uniquement avec des PPI expérimentales est, pour le moment, impossible. Un tel interactome contient à la fois des PPI fausses à cause des faux-positifs générés par les méthodes à haut débit, et à la fois, il est incomplet, car les interactions fonctionnelles transitoires sont généralement mal identifiées. De plus, tester l'ensemble des PPI de manière expérimentale est trop coûteux et prendrait trop de temps.

Ainsi, des méthodes de prédiction de PPI in silico ont été développées dans le but de permettre une exploration plus large des interactomes à moindre coût. Ces méthodes peuvent servir à guider l'identification des PPI expérimentalement ou peuvent permettre de construire un interactome plus complet contenant à la fois des PPI expérimentales et prédites, à condition que les performances de la méthode de prédiction soient correctes.

1.2.4.1 . Les réseaux de PPI

Un réseau ou graphe est une structure mathématique permettant de modéliser et de comprendre les relations entre des objets. Dans le cadre de la biologie, les réseaux permettent d'étudier de manière systémique les relations fonctionnelles entre des gènes, des espèces, des protéines... ou toute autre entité biologique. Un réseau est composé de sommets ou nœuds reliés par des arêtes qui représentent les relations liant deux sommets. Ces arêtes peuvent être non-orientées dans le cas où la relation représentée entre les deux sommets est symétrique, ou orientées dans le cas où la relation est unidirectionnelle. Dans le cadre d'un réseau de PPI, les sommets représentent des protéines, et une arête entre deux sommets représente une PPI. Les réseaux de PPI sont des réseaux non orientés, en effet il n'y a pas de notion de sens dans les interactions entre protéines, une protéine A ne peut pas interagir avec une protéine B sans que la protéine B n'interagisse avec la protéine A. La relation est symétrique.

Les réseaux sont très étudiés en mathématiques et en informatique. Ainsi, il existe tout un ensemble de métriques, de méthodes et d'algorithmes utilisables pour étudier les relations entre les sommets en se servant de la topologie du réseau. En effet, les réseaux de PPI ne sont pas construits de manière aléatoire et renferment de l'information que l'on peut extraire grâce à leur topologie. Dans un réseau de PPI, si on visualise l'ensemble des protéines participant à différents processus biologiques, on voit que les protéines participant à un même processus biologique, ont tendance à interagir plus entre elles qu'avec le reste du réseau, formant ainsi des sous-réseaux de protéines fortement interconnectées. Ainsi, grâce à des méthodes de clustering de graphes basées sur le nombre d'arêtes au sein d'un groupe relatif au nombre total d'arêtes impliquant les noeuds du groupe, il est envisageable d'inférer les processus biologiques auxquels participent des protéines de fonction inconnue [14].

Pour tirer de l'information pertinente d'un réseau, il faut que celui-ci soit le plus complet et le plus correct possible. En effet, plus le réseau construit est complet et plus les arêtes sont correctes, plus les communautés qui seront trouvées au sein du réseau seront pertinentes. Dans le cadre des réseaux de PPI, ce n'est malheureusement pas le cas. Si on prend l'exemple d'*Arabidopsis thaliana*, la plante la plus étudiée et pour laquelle il existe le plus de PPI identifiées, on ne dispose actuellement que de environ 115 000 PPI identifiées expérimentalement, toutes bases de données confondues, sachant que la taille de l'interactome a été estimée à environ 299 000 PPI. L'interactome que l'on peut construire actuellement est donc très incomplet. C'est d'autant plus vrai que la majorité des PPI identifiées par des méthodes expérimentales le sont par des méthodes à haut débit, qui ont tendance à produire de nombreux faux-

positifs et à ignorer les interactions transitoires. L'interactome que l'on peut construire aujourd'hui est donc à la fois incomplet et incorrect. Il faut donc trouver des moyens d'enrichir cet interactome en ajoutant des PPI et en arrivant à identifier les PPI incorrectes dans le réseau. Cela pourrait notamment être possible grâce aux méthodes de prédiction de PPI.

Pour avoir un réseau le plus complet possible, une stratégie possible est d'extraire un maximum de PPI de toutes les sources publiques disponibles (les bases de données de PPI) et de compléter le réseau grâce à des méthodes de prédiction de PPI.

1.2.4.2 . UniProt : la base de données universelle pour les protéines

Uniprot (Universal Protein Knowledgebase) [67] est la base de données de référence dans le domaine de la protéomique. Elle contient toutes les informations connues sur les protéine d'un très grand nombre d'organismes. Cette base est séparée en deux : (1) D'un côté il y a Swiss-Prot, une base de données où les protéines sont annotées manuellement et où les informations concernant la séquence, la fonction, la localisation cellulaire, les modifications post-traductionnelles et autres sont ajoutées manuellement par des experts. Ces annotations sont étroitement liées aux articles scientifiques qui les ont initialement présentées, garantissant une traçabilité et une fiabilité de l'information. Chaque protéine se voit attribuer un score d'annotation, allant de 1 à 5, reflétant le niveau de fiabilité de l'annotation, où une note élevée indique une bonne annotation. Swiss-Prot rassemble actuellement 570420 entrées annotées toutes espèces confondues. Chaque entrée représente une protéine et ses isoformes. (2) De l'autre côté il y a TrEMBL où les entrées sont uniquement ajoutées de manière automatisée. Chaque entrée de TrEMBL est associée à une séquence. Il y a actuellement 251 600 768 séquence dans TrEMBL.

UniProt est un hub central de l'information sur les protéines et l'identifiant associé à chaque entrée de Swiss-Prot permet généralement de faire le lien entre toutes les autres bases de données spécialisées desquelles UniProt récupère une partie des différentes informations.

1.2.4.3 . Les bases de données d'interactions protéine-protéine

Les bases de données de PPI permettent de rechercher et de télécharger des données de PPI, une étape nécessaire pour construire un réseau de PPI. Elles offrent également un accès aux méthodes et aux articles qui ont permis de les identifier. Il existe des bases de données de PPI identifiées de manière expérimentale, comme IntAct [47] et BioGRID [48], mais aussi des bases de données de PPI prédites mises en place pour stocker les résultats d'une méthode, comme la base de données prePPI [79]. Enfin, certaines bases de données mélangent les PPI expérimentales et pré-

dites, comme la base de données STRING [65].

L'International Molecular Exchange Consortium (IMEx) est une collaboration internationale majeure entre les principales bases de données de PPI [46]. L'objectif principal de l'IMEx est de centraliser les efforts d'annotation et de validation pour fournir une collection non redondante d'interactions protéine-protéine, accessible via une seule interface de recherche. L'ensemble des données de l'IMEx peut être récupéré à partir de la base de données INTACT, l'une des plus grandes bases de données publiques pour les PPI. Pour garantir une cohérence et une qualité des données élevées, des règles d'annotation et de validation communes ont été élaborées, et un registre central est utilisé pour gérer la sélection des articles à inclure dans l'ensemble de données. Malgré ces efforts, il existe toujours des défis majeurs à surmonter, notamment en ce qui concerne la redondance des données et l'accès à des informations dispersées sur plusieurs plateformes. Par exemple, l'autre grande base de données publique de PPI est BIOGRID. Elle n'est pas entièrement membre de l'IMEx. En effet, elle ne s'engage pas à fournir des données à ce consortium. Elle est considérée comme un "observateur", c'est-à-dire qu'elle est consultée en cas de décision majeure concernant les prises de décision de l'IMEx, par exemple pour les règles d'annotation et de validation.

L'IMEx est un acteur essentiel dans le domaine des PPI. Il a coordonné les stratégies d'annotation et de validation de nombreuses bases de données de PPI. Ce travail collaboratif a abouti à un ensemble de règles communes pour l'annotation et la validation des données, le développement d'un format de données commun, et à la centralisation des données. Malgré tous ces efforts de centralisation de l'information, l'ensemble des protéine participant à des PPI contenues dans INTACT ne couvre que 38,57% du protéome d'*Arabidopsis thaliana*. L'interactome de cette plante modèle est donc loin d'être complet, d'autant plus qu'une même protéine peut avoir plusieurs partenaires.

1.2.4.4 . Les méthodes de prédiction des PPI

La prédiction de PPI est ce que l'on appelle un problème de classification. C'est une méthode dont l'objectif est pour chaque paire de protéines de prédire sa classe (interaction ou pas d'interaction). Il y a plusieurs points cruciaux à prendre en compte dans le problème de prédiction des PPI.

1.2.4.5 . Le set d'apprentissage et de test

Premièrement, la construction du set d'apprentissage est essentielle car elle conditionne la prédiction. En effet, le set d'apprentissage est un ensemble de données qui sert à entraîner un modèle d'apprentissage au-

tomatique ou à fixer un seuil de décision. Il nécessite d'avoir d'une part, un ensemble de paires de protéines dont l'interaction est connue (exemples positifs), d'autre part, un ensemble de paires de protéines qui ne peuvent interagir (exemples négatifs). Si obtenir les exemples positifs est plutôt aisé (il suffit d'aller rechercher des PPI validées par plusieurs méthodes expérimentales dans les bases de données), obtenir les exemples négatifs est beaucoup plus délicat puisqu'il est rare de pouvoir affirmer que deux protéines ne vont jamais interagir car on ne connaît qu'un partie de leurs fonctions et une partie de leurs localisations cellulaires. De plus, il est important que le set d'apprentissage soit représentatif de l'ensemble des PPI et des non PPI. Pour cela, il doit contenir suffisamment de données pour couvrir le maximum de la variabilité. Les données du set d'apprentissage doivent également être de bonne qualité, c'est-à-dire exactes, complètes et cohérentes. Cette étape est essentielle et peut grandement affecter les performances des méthodes de prédiction.

Une des manières d'évaluer les performances d'une méthode de prédiction est de construire un ensemble de données de test indépendant de l'ensemble de données d'apprentissage. L'ensemble de données de test doit contenir des exemples positifs et négatifs, de manière à ce que la méthode soit confrontée à une variété de cas. Généralement, l'ensemble de données de test et d'apprentissage est construit en séparant l'ensemble des exemples positifs et négatifs de sorte que 2/3 des données soient présentes dans le jeu d'apprentissage et 1/3 des données soient présentes dans le jeu de test. Mais ces proportions peuvent être différentes si la stratégie de validation adoptée est la "cross-validation". Une fois entraînée sur le jeu d'apprentissage, la méthode de prédiction est utilisée pour classer les données du jeu de test. Les résultats de la classification sont alors comparés aux données réelles pour évaluer les performances de la méthode. On va alors pouvoir définir ce qu'est un vrai positif (VP), un faux négatif (FN), un vrai négatif (VN) et un faux positif (FP). Dans le cas de la prédiction des PPI, un VP correspond à une PPI identifiée expérimentalement et qui a été correctement prédite comme telle. Un VN correspond à une paire de protéines dont l'interaction n'a pas été identifiée expérimentalement et qui a été correctement prédite. Un FP correspond à une paire de protéines dont l'interaction n'a pas été identifiée expérimentalement mais qui a été prédite comme étant en interaction. Un FN correspond à une PPI qui a été identifiée expérimentalement mais qui a été prédite comme n'étant pas une PPI. Pour résumer, les vrais positifs et les vrais négatifs sont des instances correctement prédites, tandis que les faux positifs et les faux négatifs sont des instances incorrectement prédites.

À partir des résultats sur le jeu de test, différentes métriques sont calculées. Deux métriques complémentaires très utilisées dans le cas de la prédiction de PPI sont la précision et le rappel (Recall), appelé aussi taux de vrais positifs (TPR) ou encore sensibilité. La précision et le rappel sont des métriques complémentaires qui mesurent différents aspects des performances d'un modèle. La précision mesure la capacité du modèle à éviter les faux positifs ; plus la précision est élevée, plus la proportion de faux positifs est faible, et donc la méthode se trompe peu lorsqu'une instance est classée comme positive. En revanche, le rappel mesure la capacité du modèle à identifier les vrais positifs ; plus le rappel est élevé, plus la proportion d'instances réellement positives qui sont prédites comme positive est élevée. Les formules de ces deux métriques sont les suivantes :

$$\text{Precision} = \frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux positifs}}$$

$$\text{Recall} = \frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux négatifs}}$$

1.2.4.6 . Choisir une méthode de prédiction

Le choix de la méthode de prédiction est aussi crucial. Il existe un vaste éventail de méthodes de prédiction avec leurs avantages et leurs inconvénients. Il est important de choisir la méthode de prédiction qui convient le mieux au type de variables caractérisant l'objet à prédire, à l'objet à prédire (ici une PPI) lui même et à la dimension. Un modèle avec de bonnes performances mais qui nécessite trop de ressources peut ne pas être adapté à une application à grande échelle. À l'inverse, un modèle rapide et peu coûteux mais avec de mauvaises performances peut ne pas être adapté à une application où la précision est essentielle. Certaines méthodes requièrent plus de données que d'autres pour que l'apprentissage soit efficace. C'est le cas de la majorité des méthodes utilisant des réseaux de neurones. Un autre concept important, impacté par le choix de la méthode est l'explicabilité, c'est à dire la capacité de comprendre comment un modèle a pris la décision de classer une instance en tant que positif ou en tant que négatif. Les méthodes les plus performantes pour la prédiction de PPI, telles que les méthodes de deep learning, sont souvent des boîtes noires, ce qui signifie qu'il est difficile de comprendre

comment et pourquoi les décisions sont prises. Il peut alors être difficile pour les utilisateurs de faire confiance aux résultats de ces méthodes.

1.2.4.7 . Choisir les features

Le dernier point crucial est le choix des variables décrivant les données, que l'on appelle aussi "features" et qui vont être utilisées par la méthode pour la prédiction. Dans le cas de la prédiction de PPI, les variables couramment utilisées sont :

- Des informations extraites des séquences protéiques qui sont disponibles pour toutes les protéines identifiées. Une manière d'extraire des informations des séquences protéique est de rechercher des motifs en acides aminés ou des domaines protéiques qui ont déjà été identifiés dans des PPI. Ainsi, si on retrouve dans deux autres protéines de la même espèce ou d'une autre espèce, des motifs ou des domaines connus pour interagir l'un avec l'autre, alors il y a de grandes chances pour que les protéines interagissent effectivement l'une avec l'autre. Des bases de données spécialisées contiennent l'information sur les motifs/domaines déjà identifiés dans des PPI. On peut citer par exemple InterPro [49] ou encore ProSite [63]. Un exemple de méthode de prédiction de PPI basée sur les motifs est la méthode PIPE [54].

  Une autre manière d'utiliser les informations extraites des séquences est de créer un vecteur contenant des informations pertinentes pour la prédiction de PPI à partir de la séquence en acides aminés. Avec ce type d'approche, on ne considère pas seulement les motifs de séquences connus comme pertinents pour la prédiction des PPI. On part des séquences des protéines A et B, qui peuvent être de tailles variables, et on obtient deux vecteurs numériques de tailles fixes, qui vont refléter les caractéristiques de la séquence. Ces caractéristiques sont choisies a priori et peuvent être par exemple des propriétés physico-chimiques des acides aminés, telles que l'hydrophobicité, ou leur conservation au cours de l'évolution. Ce type d'approche est illustré dans [16], où l'Auto-covariance (AC) est utilisée pour la prédiction de PPI par des SVM (support vector machine). Dans un premier temps, les acides aminés sont transformés en valeurs numériques en fonction de différentes propriétés physico-chimique : l'hydrophobicité, l'hydrophilicité, le volume de la chaîne latérale, la charge net, l'accessibilité au solvant... Pour chaque propriété, un vecteur de la taille de la séquence est donc

généré. Pour chaque propriété physico-chimique, l'AC évalue à quel point la propriété est périodique (de période T) le long de la séquence. Les AC calculées pour des périodes allant de 1 à 30 sur chacune des deux séquences des protéines dont on veut prédire l'interaction, sont utilisées par les SVM pour prédire l'interaction.

L'homologie de séquence est un outil puissant pour prédire les PPI. Les protéines homologues sont des protéines qui ont une origine évolutive commune. Les orthologues sont un type d'homologues particulier, issus d'un ancêtre commun. Les protéines orthologues présentent des similitudes au niveau de leur séquence, en particulier au niveau des sites d'interaction. La prédiction de PPI repose sur l'idée que si deux protéines ont des orthologues qui interagissent entre eux, elles ont de fortes chances d'interagir entre elles. En effet, les interactions protéiques sont souvent conservées au cours de l'évolution. Par exemple, si on sait qu'une paire de protéines A et B interagit chez une espèce donnée, on peut prédire qu'une paire de protéines orthologues A' et B' interagira chez une autre espèce. Ces PPI conservées au cours de l'évolution sont appelées Interologues. C'est un outil puissant mais qui nécessite d'identifier des homologues chez une autre espèce pour laquelle des PPI ont été préalablement identifiées. Il existe des bases de données répertoriant des groupes de protéines orthologues, par exemple orthoDB [78] et KEGG Orthology [24]. Ce principe a été utilisé par exemple pour prédire des PPI chez *A. thaliana* à partir d'interologues récupérés chez *S. cerevisiae*, le nématode *C. elegans*, la drosophile *D. melanogaster* et l'homme [12].

L'usage des codons peut aussi être utilisé pour prédire des PPI. Un codon est une séquence de trois nucléotides, présente sur l'ADN et l'ARNm. Lorsqu'un gène est exprimé, la séquence nucléotidique du gène présent sur la molécule d'ADN est transcrite en ARNm, qui est ensuite traduite en une séquence d'acides aminés constituant ainsi une protéine. Il existe 64 codons permettant de faire la correspondance avec les 20 acides aminés protéinogènes. Les PPI peuvent être prédites en utilisant la différence d'usage des codons des deux gènes codant pour les deux protéines dont on veut prédire l'interaction[83]. L'hypothèse sous-jacente est qu'il y a un lien entre l'usage des codons et le niveau d'expression d'un gène, et que des gènes voisins ont un usage similiaire des codons. Chez certains organismes, notamment les procaryotes, les gènes voisins codent souvent pour des protéines impliquées dans la même

fonction biologique. Il y a donc plus de chance que ces protéines se retrouvent en interaction pour assurer leur fonction.

- La fonction des protéines, si elle est connue, peut être utilisée pour prédire des interactions. Le principe est que deux protéines qui participent à la même fonction biologique ont plus de chance d'interagir l'une avec l'autre. Pour prédire des interactions, les méthodes utilisent la "gene ontology" (GO) dont le but est de fournir un vocabulaire commun à la communauté scientifique pour décrire l'ensemble des fonctions des gènes et des protéines. Au sein de la GO, les termes sont organisés de manière hiérarchique du plus général au plus précis. Par exemple on pourrait trouver des termes généraux tels que "division cellulaire", qui se déclinent ensuite en termes plus spécifiques comme "mitose" ou "méiose". Cette structure hiérarchique de la GO peut être utilisée pour calculer un score de similarité entre les différents mots-clés qui décrivent les fonctions moléculaires des deux protéines d'intérêt. Ce score, basé sur la distance entre les mots clés, au sein de cette structure hiérarchique, peut ensuite être utilisé pour prédire si les deux protéines d'intérêt forment une PPI ou non. Cependant, cela nécessite d'une part, de connaître les fonctions des deux protéines dont on veut prédire la PPI, sachant qu'une protéine a généralement plusieurs fonctions. De plus, la façon dont est calculé le score de similarité peut grandement affecter les performances de ce genre d'approche. Les méthodes les plus efficaces [80] ont un score basé à la fois sur les termes en commun entre les deux protéines "au dessus" dans la structure hiérarchique de la GO (donc plutôt des termes généraux), et les termes en commun "en dessous" dans la structure hiérarchique de la GO (donc plutôt spécifiques) .

- Des variables résultant de méthodes de génomique comparative, c'est à dire, l'étude des similitudes et des différences entre les génomes. En étudiant la manière dont une paire de protéines a coévolué, il est possible de prédire une PPI. Le principe est le suivant : si deux protéines interagissent au sein d'une cellule, elles ont tendance à co-évoluer. Cela signifie que les changements dans la séquence d'une protéine entraînent des changements dans la séquence de l'autre protéine, afin qu'elles puissent continuer à interagir. En effet, si l'interface permettant la PPI est mutée sur une des deux protéines, alors pour que la structure et donc la fonction soit conservée, des mutations compensatoires peuvent être nécessaires sur l'autre protéine. C'est d'autant plus vrai si la fonction

assurée par l'interaction est essentielle à la survie de l'organisme étudié. Ces méthodes requièrent la construction d'une matrice d'alignement multiple (MSA). Une MSA est la matrice obtenue par l'alignement de plusieurs séquences protéiques de manière à maximiser la similarité ou la conservation des acides aminés. Ces MSA sont d'ailleurs utilisées dans la méthode Alphafold2 [22] et Alphafold-Multimer [11] pour prédire la structure tridimensionnelle de protéines, de complexes et également de PPI [20], voir le chapitre 5.

La génomique comparée peut aussi être utilisée pour construire l'arbre phylogénétique de chacune des protéines et d'en étudier la topologie. Des protéines qui interagissent l'une avec l'autre auront en effet un arbre phylogénétique similaire. En protéomique, un arbre phylogénétique est une représentation schématique des relations évolutives entre des protéines. Il montre comment les protéines sont liées les unes aux autres, et comment elles ont évolué à partir d'un ancêtre commun. Pour construire les arbres phylogénétique, il est nécessaire de réaliser des matrices d'alignements multiples. Des matrices de distances qui mesurent la similitude des séquences de la MSA permettent de construire l'arbre phylogénétique . Pour prédire une interaction entre une protéine A et une protéine B, il est possible de mesurer la corrélation entre les matrices de distances. Si la corrélation dépasse un seuil fixé sur un ensemble d'entraînement, la paire de protéines est prédite en interaction [50].

- La coexpression, une mesure qui permet de déterminer si deux gènes sont exprimés conjointement dans une condition donnée. En effet, pour que deux protéines interagissent, elles doivent être présentes au même moment dans la cellule. Les protéines codées par des gènes qui présentent des profils de coexpression similaires dans différentes conditions ont ainsi de plus fortes chances de former des PPI [4, 64]. Cependant, les données de coexpression dépendent des conditions expérimentales. Le fait de ne pas prédire une PPI entre deux protéines peut simplement s'expliquer par le fait que les conditions expérimentales ne permettaient pas la coexpression des gènes de ces protéines. La coexpression seule ne permet pas de prédire efficacement des PPI puisque c'est sans compter les régulations post transcritionelles et post traduction-nelle, mais c'est une information pertinente lorsqu'elle utilisée en combinaison avec d'autres.

- Les informations extraites des réseaux de PPI. Un réseau de PPI est une représentation où les protéines vont être représentées par des noeuds, et ses noeuds vont être reliés entre eux par des arrêtes. L'existence d'une arrête entre deux noeuds signifie qu'une PPI a été identifiée entre deux protéines. Cette représentation permet l'étude des PPI à l'échelle d'un processus biologique voir d'un interactome complet. La topologie d'un réseau de PPI peut être utilisée pour prédire d'autres PPI, par exemple en utilisant une structure particulière des réseaux : la clique. Une clique est un ensemble de noeuds qui sont tous connectés entre eux. Dans un réseau de PPI, c'est le cas lorsqu'un ensemble de protéines sont toutes des sous unités d'un même complexe. Si l'ajout d'arêtes entre les noeuds d'une clique et une protéine connectée à certains noeuds de cette clique permet d'obtenir une clique plus grande, alors cela pourrait signifier que la protéine ajoutée à la clique est elle aussi une sous unité du même complexe [77].

- La structure 3d des protéines. Il existe deux types d'approches.

  La première est de rechercher dans les structures des protéines dont on veut prédire l'interaction des motifs similaires présents au niveau des interfaces des protéines de complexes connus. C'est ce que fait la méthode PRISM [70]. Un point intéressant soulevé par cette méthode est qu'elle arrive parfois à prédire correctement des PPI même si les structures des protéines d'intérêt sont globalement très différentes de celles des sous unités des complexes connus. Cela met en avant que des protéines très différentes peuvent parfois interagir via des motifs structuraux similaires à l'interface entre les sous-unités.

  Une autre approche utilisant les structures est d'utiliser des méthodes de docking. Les méthodes de docking sont des méthodes qui servent à trouver les positions et les orientations optimales (en terme d'énergie) d'une structure protéique par rapport à une autre, de manière à modéliser la structure du complexe qu'elles peuvent former. Pour cela, elles utilisent notamment une énergie d'interaction calculée à partir des propriétés physico-chimiques présentent à la surface des protéines afin de trouver un ensemble d'interfaces protéiques potentielles minimisant cette énergie. Ce ne sont donc pas des méthodes qui ont pour but de prédire des couples de protéines en interaction. Cependant une interface très favorable à l'interaction entre deux protéines pourrait signifier qu'il existe bien une interaction entre elles. Pour prédire des PPI,

certaines méthodes utilisent la distribution des valeurs d'énergie d'interaction générées par une méthode de docking pour un complexe donné. En effet, les protéines en interaction présentent des distributions d'énergies significativement différentes de celles n'en formant pas [75]. L'inconvénient majeur de ces méthodes est qu'elle nécessite de connaître la structure des protéines dont on veut prédire l'interaction. C'est loin d'être le cas chez la plupart des organismes. Cependant, les avancés majeures réalisées en terme de prédiction de structures 3D, notamment avec Alphafold2 [22], offrent de nouvelles opportunités pour les méthodes utilisant l'information structurale. La base de données alphafoldDB a mis à disposition environ 200 millions de structures de protéines couvrant environ 1 million d'organismes. L'utilisation de ces modèles nécessite de bien prendre en compte les mesures de qualité de prédiction fournies par la méthode Alphafold2 pour n'utiliser que des modèles fiables.

Certaines méthodes utilisent plusieurs des features présentées précédemment. Varier la nature des données utilisées pour la prédiction des PPI peut permettre d'améliorer les performances des méthodes de prédiction car les variables utilisées peuvent fournir des informations complémentaires. Par exemple, la méthode PrePPI est une méthode qui utilise de l'information structurale ainsi que des données de co-expression, des données sur la fonction des protéines et un profil phylogénétique. Lors de la publication de l'article en 2012, PrePPI [79] a démontré de meilleurs performance que ses concurrents de l'époque dans les cas où l'information structurale était disponible. Comme on le verra plus tard, Alphafold2 combine deux type d'information différentes.

2 - Mon travail de thèse

L'objectif initial de ma thèse était de construire un réseau de PPIs d'*Arabidopsis thaliana* qui soit le plus complet et le plus exact possible, afin d'améliorer les connaissances des processus biologiques de cette plante modèle.

Une première partie de ma thèse a donc consisté à développer un package R permettant de construire et d'analyser un réseau de PPIs, à partir des données issues des différentes bases de données publiques mais aussi de données privées (PPIs identifiées expérimentalement au laboratoire). J'ai ainsi développé le package APPINetwork que j'ai présenté au cours de divers événements et congrès. Ce travail m'a aussi valu une publication en tant que premier auteur [14].

Puis, pour compléter le réseau de PPIs et corriger les éventuelles fausses interactions, nous souhaitions initialement développer une méthode de prédiction de PPI s'appuyant notamment sur des cartes en deux dimensions représentant différentes propriétés physico-chimiques de la surface des protéines (stickiness, hydrophobicité, variance circulaire, etc.) [60], l'énergie d'interaction obtenue par docking moléculaire [74] à la surface de chacune des protéines et leur propension à l'interaction (cartes IPOPS [59]). Afin d'obtenir les cartes de propension à l'interaction, je devais réaliser le docking des protéines dont la structure était disponible chez *A. thaliana*. Comme ces dockings nécessitaient du temps pour être obtenus, nous devions travailler en parallèle sur le développement de la méthode de prédiction de PPI chez S. cerevisiae. En effet, nous disposions déjà des données de docking croisées chez cet organisme, qui avaient été générées par un doctorant précédent.

Ainsi, durant le début de ma thèse, j'ai dû réaliser un inventaire de ces dockings, ce qui représentait 12 To de données. Une fois cet inventaire réalisé, j'ai écrit un pipeline permettant la génération des cartes IPOPS pour l'ensemble des 471 protéines pour lesquelles nous possédions des résultats de dockings croisés. Pour ces protéines, j'ai aussi généré les cartes de propriétés physico-chimiques grâce au logiciel SURMAP auquel j'ai apporté quelques modifications qui m'ont valu une publication quatrième auteur [60]. Pour la suite, si l'arrivée d'AlphaFold2 a été une véritable révolution dans le domaine de la prédiction des structures des protéines, l'impact s'est fait ressentir également dans le domaine de la prédiction des PPI et a par conséquent fortement impacté ma thèse. En effet, des articles tels que celui de Humphrey et al [20] ont montré qu'en utilisant certaines métriques produites par AlphaFold2, il est possible de prédire des PPI avec un niveau de confiance élevé. Nous avons donc décidé de

mettre en place une méthode de prédiction de PPI utilisant les résultats fournis par d'AlphaFold2 dans sa version multimer et d'analyser les résultats de nos prédictions sur la base des différentes cartes que j'avais générées.

Afin de pouvoir prédire un grand nombre de PPI, j'ai donc tout d'abord mis en place la stratégie du papier [20] qui consistait à utiliser une méthode très rapide mais moins performante qu'AlphaFold2 pour présélectionner les paires de protéines les plus probables et prédire ainsi avec AlphaFold2 un nombre beaucoup plus réduit de complexes. L'approche que j'ai ainsi développée utilise la version multimer d'AlphaFold2 [11] qui venait d'être publiée. Elle permet d'extraire des résultats d'AlphaFold2-multimer des probabilités d'interactions que l'on peut ensuite utiliser pour prédire des PPI mais aussi comme critère de fiabilité des PPI. J'ai ensuite testé cette méthode sur le jeu de PPI de *S. cerevisiae* que m'a fourni Humphrey pour évaluer le gain de performance de prédiction. Mais je n'ai pu comparer mes résultats à ceux publiés car je ne disposais pas de leur jeu de non-interactions (couples de protéines choisis aléatoirement) ni des ressources de calcul nécessaires pour réaliser les prédictions sur autant de couples aléatoires que les leurs.

J'ai alors découvert qu'une version optimisée d'AlphaFold2, ColabFold [40] venait d'être publiée. Je me suis alors demandé si l'étape de présélection des paires de protéines les plus probables, mise en place précédemment était encore nécessaire. J'ai alors mis au point une stratégie de prédiction utilisant AlphaFold-multimer dans sa version ColabFold, sans passer par l'étape de présélection des PPI. J'ai modifié plusieurs paramètres d'AlphaFold-multimer de façon à obtenir des temps de calcul réduits et j'ai comparé les résultats des deux stratégies sur le même jeu de données de PPI et de paires de protéines aléatoires de *S. cerevisiae*.

Les résultats obtenus chez *S. cerevisiae* étant positifs, j'ai ensuite appliqué la meilleure stratégie à la prédiction des PPI d'*A. thaliana* sur un jeu de PPI déterminées par plusieurs méthodes expérimentales (comme chez *S. cerevisiae*) et sur des couples de protéines aléatoires.

3 - APPINetwork : un package R pour la construction et l'analyse de réseau de PPI

3.1 . L'origine d'APPINetwork

Le sujet de ma thèse étant l'annotation fonctionnelle du protéome d'*A. thaliana* via l'analyse de son interactome, j'ai débuté ma thèse par la construction de cet interactome. J'ai donc commencé par extraire les PPI de cette espèce dans les bases de données de PPIs. Je me suis alors rendu compte qu'un certain nombre d'entre elles étaient présentes dans une base de données mais pas dans l'autre et qu'il était donc intéressant d'en interroger le plus possible pour avoir une liste de PPI la plus exhaustive possible. Mais le problème que j'ai rencontré lorsque j'ai voulu regrouper les PPIs provenant de différentes base de données, c'est que les identifiants protéiques n'était pas toujours les mêmes d'une base de données à l'autre.

3.1.1 . Uniformisation des indentifiants

Il m'a donc fallu dans un premier temps uniformiser les identifiants protéiques avant de pouvoir éliminer la redondance entre des différentes bases. Pour faire cela, j'ai du créer tout d'abord un fichier de correspondances entre les différents identifiants de chaque protéine du protéome, à partir des informations extraites de la base de données Uniprot. J'ai rencontré de multiples problèmes avec parfois plusieurs protéines pour un même gène, parfois plusieurs gènes pour une même protéine ...etc. Il a donc été long et laborieux d'établir ce fichier de correspondances, mais une fois établi, il a grandement facilité la construction de l'interactome .

3.1.2 . Elimination de la redondance

Sur l'interactome ainsi construit, j'ai éliminé la redondance entre les PPIs présentes plusieurs fois dans le fichier soit parcequ'elles se trouvaient dans plusieurs bases de données soient parcequ'elles avaient été identifiées par différentes expériences et donc publiées dans différents articles. Mais lors de cette étape, j'ai conservé l'origine de chaque PPI, c'est-à-dire la liste des bases de données d'où elle provient et surtout la liste des identifiants Pubmed des articles où elle a été identifiée expérimentalement. Le nombre de publications associées à chaque PPI est un indicateur de la fiabilité de l'interaction. C'est donc une information utile, car comme je l'ai indiqué précédemment. Certaines technologies à haut

débit fournissent beaucoup de faux positifs.

3.1.3 . Réseaux d'ordre 1 et réseaux d'ordre 2

Les PPI de l'interactome sont des interactions d'ordre 1, c'est à dire que les protéines sont liées directement l'une à l'autre. Mais parfois, l'analyse d'un réseau de PPI nécessite d'étendre le réseau à des protéines liées indirectement aux protéines d'intérêt. C'est le cas lorsqu'on cherche à annoter le protéome d'une espèce à partir de son interactome. En effet, les protéines impliquées dans un même processus biologiques interagissent fortement les unes avec les autres et forment donc des des sous réseaux de protéines fortement inter-connectées. J'ai donc construit l'interactome d'ordre 2, c'est à dire le réseau formé par les interactions médiées par une protéine tierce. Mais le réseau d'ordre 2 que je construis à partir d'une liste de protéines d'intérêt est un peu particulier. En effet, pour ne pas avoir dans les sous réseaux, des protéines qui n'ont rien à voir avec le processus biologique d'intérêt, je ne conserve que les PPI d'ordre 2 dont les deux protéines sont liées à des protéines en interaction avec des protéines d'intérêt (pour d'avantage d'explication voir le papier joint à ce chapitre).

3.1.4 . Analyse du réseau d'interaction protéine-protéine

Une fois l'interactome construit, il fallait l'analyser afin d'annoter le protéome. Pour rechercher des sous réseaux de protéines fortement inter-connectées dans l'interactome d'ordre 2, j'ai utilisé le logiciel Tfit [51], un outil fondé sur un algorithme de clustering de graphe basé sur la modularité, développé par des collaborateurs de l'INRIA et utilisé par l'équipe dans plusieurs projets.

3.1.5 . Mise à jour régulière

Certaines bases de données de PPI étant mises à jour très régulièrement (tous les mois pour certaines), j'ai dû moi aussi faire de nombreuses mises à jour du fichier de correspondances et des fichier de PPI.

3.1.6 . Développement du package APPINetwork

Pour faire tout ce que je viens de présenter, j'avais écrit de nombreux script R. Et en présentant mon travail au biologiste de mon laboratoire, je me suis aperçu qu'ils étaient intéressés par ce que j'avais développé. En effet, la seule chose qu'ils utilisaient pour construire et visualiser le réseau d'interactions protéiques impliquant leurs protéines intérêt est le logiciel STRING [65], un outil certes pratique et convivial, mais dont la base de données est très incomplète et qui fournit des informations sur les protéines qui sont souvent erronées. Nous avons donc décidé de développer un nouveau package R avec une interface conviviale permettant à n'importe quel utilisateur de construire et d'analyser le réseau d'ordre 1 ou

d'ordre 2 impliquant ses protéines d'intérêt. Nous avons nommé ce package APPInetwork (Analysis of Protéine Protéine Interaction network) et nous l'avons publié il y a un an dans la revue PeerJ. Il est par ailleurs téléchargeable sur https ://forgemia.inra.fr/GNet/appinetwork et il est accompagné d'un guide utilisateur et de plusieurs exemples d'utilisation.

Dans ce chapitre, je ne m'étendrai pas davantage sur APPINetwork car il est décrit avec précision dans l'article ci-joint. Les résultats de la comparaison de ses performances avec celle des Package existants y sont par ailleurs fournis.

3.2 . L'article de référence d'APPINetwork

APPINetwork: an R package for building and computational analysis of protein–protein interaction networks

Simon Gosset[1,2,*], Annie Glatigny[3,*], Mélina Gallopin[3], Zhou Yi[3], Marion Salé[3] and Marie-Hélène Mucchielli-Giorgi[1,2]

[1] Université Paris-Saclay, CNRS, INRAE, Université Evry, Institute of Plant Sciences Paris-Saclay (IPS2), Gif-sur-Yvette, France
[2] Université de Paris, Institute of Plant Sciences Paris-Saclay (IPS2), Gif-sur-Yvette, France
[3] Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), Gif-sur-Yvette, France
[*] These authors contributed equally to this work.

ABSTRACT

Background. Protein–protein interactions (PPIs) are essential to almost every process in a cell. Analysis of PPI networks gives insights into the functional relationships among proteins and may reveal important hub proteins and sub-networks corresponding to functional modules. Several good tools have been developed for PPI network analysis but they have certain limitations. Most tools are suited for studying PPI in only a small number of model species, and do not allow second-order networks to be built, or offer relevant functions for their analysis. To overcome these limitations, we have developed APPINetwork (Analysis of Protein–protein Interaction Networks). The aim was to produce a generic and user-friendly package for building and analyzing a PPI network involving proteins of interest from any species as long they are stored in a database.

Methods. APPINetwork is an open-source R package. It can be downloaded and installed on the collaborative development platform GitLab (https://forgemia.inra.fr/GNet/appinetwork). A graphical user interface facilitates its use. Graphical windows, buttons, and scroll bars allow the user to select or enter an organism name, choose data files and network parameters or methods dedicated to network analysis. All functions are implemented in R, except for the script identifying all proteins involved in the same biological process (developed in C) and the scripts formatting the BioGRID data file and generating the IDs correspondence file (implemented in Python 3). PPI information comes from private resources or different public databases (such as IntAct, BioGRID, and iRefIndex). The package can be deployed on Linux and macOS operating systems (OS). Deployment on Windows is possible but it requires the prior installation of Rtools and Python 3.

Results. APPINetwork allows the user to build a PPI network from selected public databases and add their own PPI data. In this network, the proteins have unique identifiers resulting from the standardization of the different identifiers specific to each database. In addition to the construction of the first-order network, APPINetwork offers the possibility of building a second-order network centered on the proteins of interest (proteins known for their role in the biological process studied or subunits of a complex protein) and provides the number and type of experiments that have highlighted each PPI, as well as references to articles containing experimental evidence.

Submitted 27 October 2021
Accepted 19 September 2022
Published 4 November 2022

Corresponding author
Simon Gosset,
simon.gosset1@universite-paris-saclay.fr

Academic editor
Kenta Nakai

Additional Information and Declarations can be found on page 12

DOI **10.7717/peerj.14204**

How to cite this article Gosset S, Glatigny A, Gallopin M, Yi Z, Salé M, Mucchielli-Giorgi M-H. 2022. APPINetwork: an R package for building and computational analysis of protein–protein interaction networks. *PeerJ* **10**:e14204 http://doi.org/10.7717/peerj.14204

Conclusion. More than a tool for PPI network building, APPINetwork enables the analysis of the resultant network, by searching either for the community of proteins involved in the same biological process or for the assembly intermediates of a protein complex. Results of these analyses are provided in easily exportable files. Examples files and a user manual describing each step of the process come with the package.

Subjects Bioinformatics, Molecular Biology
Keywords Network clustering, Protein–protein interaction, Network, Protein complex intermediaries

INTRODUCTION

Protein–protein interactions (PPIs) are central to many cellular processes. PPIs are identified and characterized experimentally by different methods which determine whether two proteins make physical contact or if they belong to a transient or permanent complex. There are advantages and limitations to any method of identifying and measuring PPIs, reviewed by *Snider et al., 2015*. Well-known drawbacks of some methods are the identification of proteins that interact in the experimental conditions but not in a biological context (false positives) or failing to identify known or probable interactions that are biologically significant (false negatives). To fully appreciate the range of PPIs that are possible within the predicted proteomes of several model organisms (*Tran, Hamp & Rost, 2018*), it is of interest to supplement the information on experimentally identified PPIs with PPI predictions (*Humphreys et al., 2021*).

PPI data is stored in repositories of various formats. The experimental results or computing methods used to identify or predict PPIs are diverse. In addition, the IDs and descriptions are not comparable from one database to another. To ensure easy access to the data and reliable outputs, the Human Proteome Organization (HUPO) initiative (*Orchard & Hermjakob, 2008*) and the International Molecular Consortium (IMEx) (*Porras et al., 2020*) have defined guidelines including accepted terminology and standardized data formats that should be used by authors reporting PPIs. Many curators have already adopted these principles for handling the data which is greatly facilitating exchanges and comparisons, although some disharmony still exists.

The Universal Protein Resource UniProt (*Apweiler et al., 2004*; *The UniProt Consortium, 2018*) is a collection of sequences with functional annotations and diverse information about each protein. The nomenclature and vocabulary are standardized, and various formats are available. This rich and user-friendly resource provides the reference proteome of species and several feature viewers that summarize and give access to data on localization, interactions, and molecular structures.

There is a large variety of biomolecular interaction databases. Some are specific to a particular type of interaction, others focus on a given organism type (fly, yeast, bacteria), or disease (*Miryala, Anbarasu & Ramaiah, 2018*). In this article, we will only consider some of the most frequently used PPI databases. The BioGRID database of physical, genetic, and chemical interactions reported in various organisms is updated monthly (*Oughtred et al.,*

2019; Oughtred et al., 2021). The data can be downloaded in multiple formats, and more tools and resources are provided for analysis. The iRefIndex database (Razick, Magklaras & Donaldson, 2008) is a secondary database that collates non-redundant data on interactions from freely available sources. A confidence score is calculated for each accession. The open-source IntAct database provides interaction data derived from literature curation or direct submission as well as interactomes from different species or datasets. APID (Alonso-López et al., 2019) provides curated interactomes of 400 organisms based on PPI information from six primary databases of molecular interactions and experimentally resolved 3D structures. APID also includes a data visualization tool. APID's user-friendly and intuitive interface can be used to look for physical, genetic, or predicted interactions alongside expression or localization data from an input set of genes of interest. The Proteomics Standard Initiative Common QUery InterfaCe (PSICQUIC) (Aranda et al., 2011) aggregates molecular interaction data from 23 servers. In the first PSICQUIC version, each PPI was described by 15 fields corresponding to PSI-MITAB2.5 format. Since this first version, other file formats give more information about the reported interactions and more facilities to the users. STRING, one of the most popular tools for representing PPI networks (Szklarczyk et al., 2019), aggregates details of experimental or predicted physical and functional interactions from other databases. In total, STRING provides protein interaction data with associated confidence scores from 5090 organisms. The network resulting from the user's request can be easily exported in different formats of text and image files.

The analysis of protein interaction networks (PIN) is of great interest when studying biological activities, pathways, or drug targeting and is the reason why many web tools or plugins for visualization and analysis of protein interaction networks have been developed. For example, Pathguide (Bader, Cary & Sander, 2006) provides a list and brief description of 702 pathways and molecular interaction resources.

Cytoscape (Shannon et al., 2003) plugins have been developed to export PPI data and visualize PPI networks (Martin et al., 2010; Doncheva et al., 2019; Holmås et al., 2019; Legeay et al., 2020). Here, we will only present three of them because their functionalities are close to those of the package we developed. GeneMANIA (Warde-Farley et al., 2010) imports an interaction network from a list of genes with their annotations and putative functions. The interactions of the network are associations, *i.e.*, the most closely related genes to a query gene set are identified using guilt-by-association. With this approach, new members of a pathway or a complex are found and weights are assigned to the interactions. From a combination of the most trusted datasets from UniProt, Intact, and other curated sources, BioGateway (Antezana et al., 2009) provides a network of interactions of different types, among which are PPIs annotated with GO terms. The interactome browser mentha (Calderone, Castagnoli & Cesareni, 2013) provides interactomes of eight model species based on PPI data from databases set up by the IMEx consortium.

Other related tools have been developed in the R programming language (Ihaka & Gentleman, 1996) to display the shortest paths of functional interaction between proteins and are provided by Bioconductor (Gentleman et al., 2004). The package Path2PPI (Philipp, Osiewacz & Koch, 2016) helps researchers find proteins and interactions of pathways or

biological processes in fully sequenced organisms for which virtually no PPI is known. With the cisPath package (*Wang et al., 2015*), cloud users can integrate downloaded functional information on PPI from different online databases or private data to construct, visualize, manage, and share functional protein interaction networks.

In developing these tools, the respective authors carefully considered how to benefit the most from existing data when seeking to answer different biological questions. Depending on the topic, one tool may be better than another. In their article from *2016*, *Pan et al.* reviewed the computational approaches to analyze PINs. While the above mentioned tools successfully integrate information on first-order neighbors in the network, they do not readily deal with all the experimental second-order PPIs involved in the biological process of interest. However, to find clusters in a PPI network, it is necessary to account for the second-order PPIs.

In the present article, we describe APPINetwork, an R package for constructing PPI networks to search for (1) sets of proteins involved in the same biological process and (2) proteins or protein sub-complexes that play a role in the assembly of a protein complex (*Glatigny et al., 2017*). Starting from an input set of proteins, APPINetwork builds the PPI networks of the first or second-order, by using PPIs derived from all the available PPI databases and potentially any privately held data. APPINetwork thus provides the most exhaustive possible network of PPIs, whether experimental or predicted. The fact that this network integrates public and private data, makes the package particularly useful for all research teams who have identified new PPIs, and for proteomics platform groups that have accumulated large datasets of unpublished PPIs. Through a user-friendly interface, APPINetwork allows users (1) to choose the specie the user is studying from among hundred species currently included, (2) to select the queried PPI databases including any proprietary data files, (3) to select the order of network desired (first or second order), and (4) select the analysis to perform. We first present how the package is implemented and then discuss the advantages of similar tools.

MATERIALS AND METHODS

Implementation

All functions of the package APPINetwork are implemented in R except for the script to search for all proteins involved in a biological process (see Fig. 1) which was developed in C (*Gambette & Guénoche, 2011*), and the scripts that format the BioGRID data file (see Fig. 1) and generate the ID correspondence file (see Fig. 1) which were implemented in Python. Indeed, the use of Python is optimal for writing functions for text mining large text files such as BioGRID files (*Oughtred et al., 2021*).

Minimal configuration and dependencies

The APPINetwork package can be deployed on Linux and macOS operating systems (OS). It can also be deployed on Windows with prior installation of Rtools and Python 3 (see the "readme" file). Minimal requirements are 64 bits Unix-based OS (Linux/macOS) or the 64-bit version of Windows.

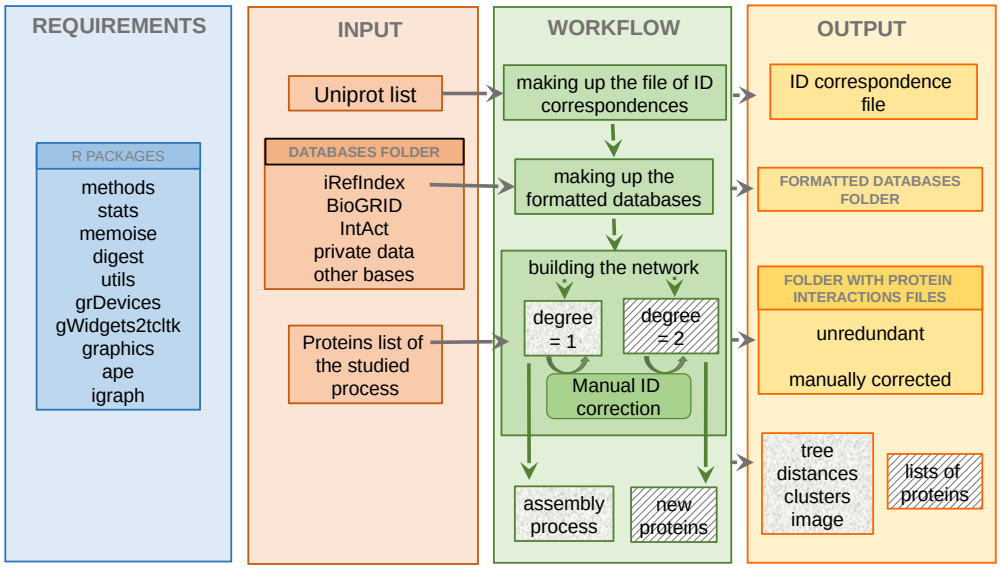


Figure 1 Overview of the APPINetwork package. Illustration of requirements R packages (blue section), inputs (orange section): databases, UniProt file text of the studied organism and lists of proteins of interest, workflow (green section) and outputs (yellow section): all files that APPINetwork provides to the user. The "new proteins" in the green section are proteins newly identified by APPINetwork as playing a role in the biological process of interest. The "lists of proteins" in the yellow section are the lists of all proteins that make up each sub-network potentially associated with a biological process.

Full-size DOI: 10.7717/peerj.14204/fig-1

Installation

APPINetwork is an R package and requires R version 3.2.0 or later versions. The package can be downloaded and installed from GitLab, directly from the R console using the R command *devtools::install_gitlab("Gnet/appinetwork", host = "https://forgemia.inra.fr")*. Its installation thus requires the devtools package (*Wickham, Hester & Chang, 2019*). An installation tutorial can be found in the project repository (https://forgemia.inra.fr/GNet/appinetwork). All the R packages required in APPINetwork are automatically installed.

Graphical interface

APPINetwork has a graphical user interface for users who are less familiar with using command lines. This graphical interface is based on the gwidgets package (Verzani, 2019. Graphical windows, buttons, and scroll bars allow the user to select or enter an organism name, select files, and choose network parameters or specific methods for network analysis.

Required data files

To create a network of PPIs involving proteins known for their role in the biological process of interest, APPINetwork requires different flat files. These must be prepared or downloaded beforehand.

The first file (named "input list") contains different names or IDs (Name, UniProtID, UniProtName, alias, and Systematic Name) of the proteins involved in the biological process. To prepare this file, the user must adhere to the format presented in the user

guide that comes with the APPINetwork package, available in the GitLab repository. The second file required by APPINetwork to standardize protein IDs between PPI databases is the UniProt file (in .txt format) of the proteome of the organism to study. It can be downloaded from the UniProt website (https://www.uniprot.org/, see section "Download the UniProt file" of the user guide).

The other files to download are PPI files from the databases iRefIndex, IntAct, BioGRID, and any other private or public databases chosen by the user. The package enables automatic formatting and updating of the IrefIndex, IntAct, and BioGRID databases. If the user wants to integrate other databases or personal data into the network, the files must be formatted independently before use. The PPI files should contain 15 columns as follows: UniProt identifiers for each protein (*uid* and *alias*), identification method, author of the publication, PubMed IDs, taxon name, interaction type (physical or genetic), name of the databases, and the name of the gene encoding each protein. The format is described in the user guide.

The user guide describes all the formats of the different files needed at each step. By way of illustration, some example files are provided with the user guide. The user can use them to practice using APPINetwork.

Parameters of PPI Network

Different types of networks should be built, depending on the kind of analysis to be performed. For example, to search for all proteins involved in the same biological process, the user should search for dense clusters in a network with second-order PPIs determined by physical or genetic methods ensuring there are no self-loops that may impact the clustering (see Discussion). On the contrary, if searching for assembly intermediates of a protein complex, the PPI network should be of the first-order and composed of PPIs determined by physical experiments or predicted from structural information with self-loops to account for any dimers.

APPINetwork thus offers different options to build the network before analyzing it. These options are (i) the physical or genetic experimental method used for detecting the PPIs, first or second-order PPIs, (ii) the removal of all proteins involved in only one PPI or not, (iii) the removal of proteins involved in only one second-order PPI or not, (iv) the removal of self-loops or not. Our second-order PPIs are particular because they involve two proteins that interact by two different pathways with the proteins of the input list. They thus facilitate the search for small clusters (*Glatigny et al., 2011*. Even with this method of constructing second-order interactions, there may be many proteins in the second-order network that are not specific to the biological system under study. This is the case when a protein in the first-order network interacts with more than a hundred proteins. To work around this issue, once the choice is made to build a second-order network, APPINetwork offers an option that filters out proteins if they interact with a number of proteins that exceeds a threshold fixed by the user.

APPINetwork analysis tools

The APPINetwork package offers two very different analysis tools. (1) The first tool can be used to search for assembly intermediates of a protein complex (*Glatigny et al., 2017*). The

underlying hypothesis is that proteins belonging to an assembly intermediates interact with the same proteins and thus have more common partners than the other subunits of the complex. Consequently, the subunits of a protein complex (the proteins constituting the final complex) are aggregated according to the number of partners they have in common. The resulting clusters are assembly intermediates. (2) The second tool is designed to search for all the proteins involved in the biological process of interest, by searching for clusters of proteins that are strongly interconnected but weakly connected to the rest of the network (*Gambette & Guénoche, 2011*).

RESULTS

To start with APPINetwork

The graphical interface of the APPINetwork menu offers the choice between five actions: (i) construct a correspondence file between different IDs; (ii) format iRefIndex, IntAct, or BioGRID PPI files; (iii) build a network; (iv) identify proteins involved in a biological process, and (v) identify the assembly intermediates of a protein complex (see Fig. 2 and the user guide). When using APPINetwork for the first time, actions (i), (ii), and (iii) must be executed successively. Indeed, formatted PPI files are required to build the network for which a correspondence file with the different identifiers is necessary. On the second use, if the user is continuing to work on the same organism, it is not necessary to execute steps 1 and 2 again. Another network can be built from another input list or from the same input list but with different parameters. In the same way, any network built with another tool, if formatted as described in the user guide, can be analyzed with APPINetwork using actions (iv) or (v).

Package functionalities

Making up the correspondence file of IDs

To build a correspondence file of IDs, the user must choose one of eight organisms from the drop-down menu. Selecting the "other" option opens a second window where the name of the organism of interest can be typed in. Then, the user must select the UniProt file of the species previously saved on the computer (see Fig. 2 and the user guide). The result of this action is a correspondence file of IDs that stores all names and IDs (Gene Name, RefSeq number, Protein Name, Gene ID, BioGRID ID, UniProt IDs) of each protein of the studied proteome.

Updating and formatting of the databases

To format the PPI files previously downloaded from iRefIndex (*Razick, Magklaras & Donaldson, 2008*), IntAct (*Orchard et al., 2014*; *Del Toro et al., 2022*) and BioGRID (*Oughtred et al., 2021*) databases, the user must choose the name of the database. A window corresponding to his/her choice is then displayed, allowing the user to choose the name of the organism and the file to format. The iRefIndex file (*Razick, Magklaras & Donaldson, 2008*) is split into different files, each containing PPIs from a single initial database. To format a BioGRID file, the user must choose whether to keep the PPIs of putative proteins and then select the UniProt file of the organism of interest (see Fig. 2 and

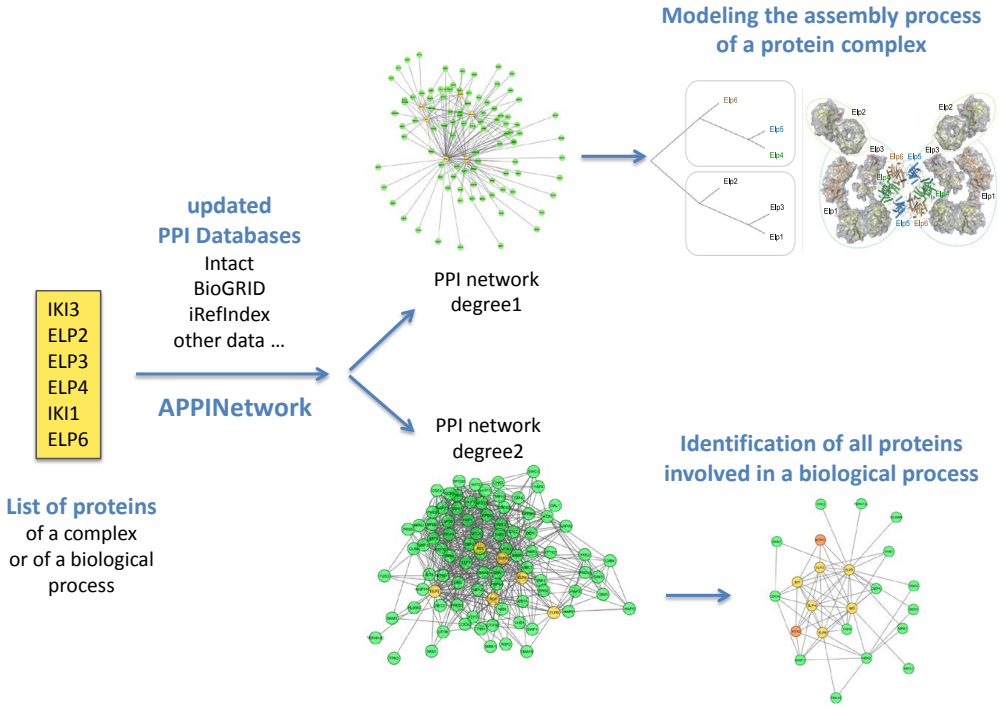


Figure 2 Outline of analysis types of networks obtained with APPINetwork for the ELP complex of *Saccharomyces cerevisiae*. With a list of the six proteins of the elongation factor of Saccharomyces cerevisiae (yellow box), the user can either build a first order network to search for assembly intermediaries (upper part), or a second order network to search for all the proteins interacting with the six proteins. To do this, he/she can use the TFit clustering algorithm.

Full-size DOI: 10.7717/peerj.14204/fig-2

the user guide). The resulting formatted files have 15 columns (see the user guide). They contain only interactions between two proteins of the same strain of the studied species.

Building of the network

To build a network (see Fig. 3), the user must select the "input list" file that has been prepared in advance (see section "required files" in the section "Material and Methods"). Then the formatted PPI files are selected by clicking the "select database" button. The user must also select an ID correspondence file by clicking on the "ID correspondence file" button and choosing the organism. Finally, the user must decide whether the network should contain (a) PPIs determined experimentally by physical or genetic methods or both, (b) PPIs of the first or second order, (c) proteins interacting (whether first or second order PPI) with a single protein of the input list (termed "unique link"), (d) interactions of a protein with itself (see Fig. 2 and the user guide), and if necessary (e) indicate a maximum number of first-order protein partners.

The script looks for all the PPIs involving proteins of the input list (first-order PPI) inside the formatted files of PPIs. In case of discrepancy between gene names or protein IDs, the program sends a warning to the user, who can then manually correct them. It removes redundant PPIs and records the IDs of publications mentioning each PPI. It

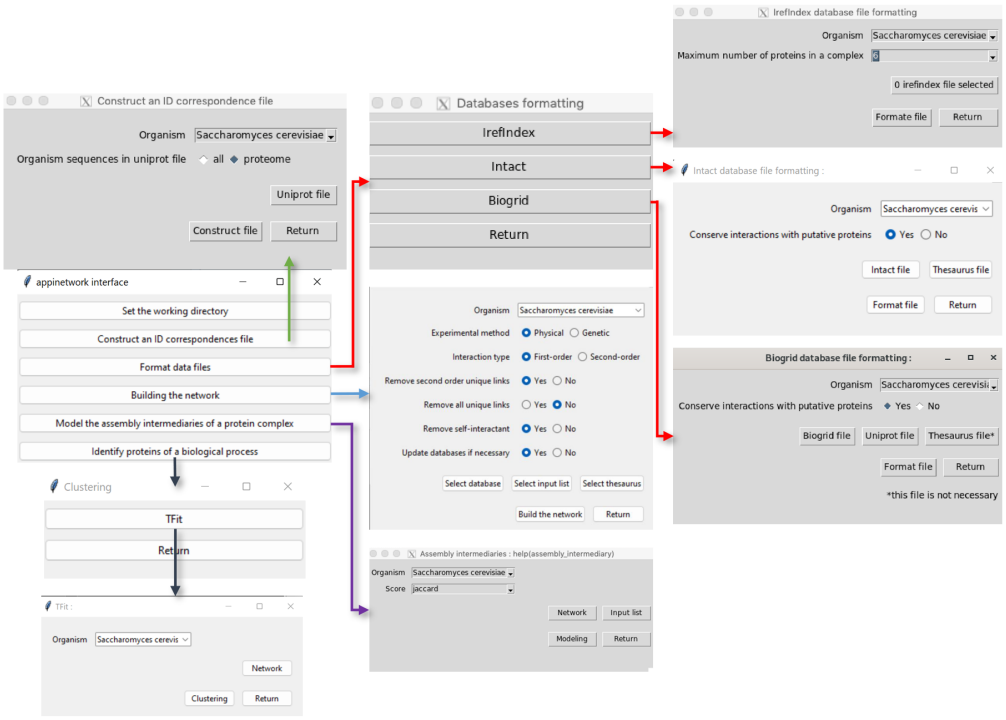


Figure 3 Procedure to use APPINetwork. Graphical interfaces allow the user to build and analyze a network. With APPINetwork the user can construct an ID correspondence file (green arrow); can format databases of his/her choice (red arrows); can build a network (light blue arrow). The user has to choose the parameters he/she wants to use by clicking on the interface, then he/she can analyze the network. To study the assembly process from a first order network, he/she has to choose one the six similarity scores ; from a second-order network and to study functional interactions (dark blue arrow) he/she can use TFit.
Full-size DOI: 10.7717/peerj.14204/fig-3

calculates the number of these publications and records it in the file because it is an index of the reliability of the interaction. Other information related to each PPI, such as the type of experiment, the biological function of each network protein the name of the organism is also stored in the file. The network of order one or two is saved as a flat file with 13 columns containing all PPIs of order one or two and related information (see Materials and Methods), that can be exported to other analysis tools.

APPINetwork thus gives a lot of information on PPIs and offers to build a particular second-order network, which other software does not offer. The downside is that it takes time. As an example of the time required, on a laptop computer with 32 GB of memory and an Intel Core I7, the computation time was about 7 min for a network built from a database of 683,389 PPIs, with a threshold of 300 for the maximum number of partners of each protein.

Analysis of the network

To identify proteins involved in a biological process, the user has only to choose a second-order PPI network (see Fig. 2 and the user guide). The clustering method TFit then identifies

small clusters of highly interconnected proteins containing proteins from the input list and other proteins potentially involved in the biological process of interest that are good candidates for validation. Results of the clustering with TFit are provided in a flat file (with the extension .clas), where clusters are numbered and are provided as a list of proteins separated by semicolons.

To identify assembly intermediates of a protein complex, the user should select a first-order network, the input list and to click on "modeling" to build the assembly model. Finally, the user should choose the metric used to model the assembly of the complex that is described in (*Glatigny et al., 2017*) (see Fig. 2 and the user guide). Results are provided in different files: (1) a text file ("score_distance_matrix.txt") with a matrix of the distance values between the subunits; (2) a text file ("hc.txt") showing how the subunits are aggregated; (3) a jpeg picture of the hierarchical tree ("tree.jpeg"); and (4) text files for each subcomplex ("Proteins_subcomplex_name.txt"), containing all proteins interacting with proteins of the subcomplex.

The computation time for both TFit and the identification of assembly intermediates is instantaneous.

DISCUSSION

APPINetwork is more than a PPI network building tool since it also offers two clustering methods to analyze the resulting PIN. It can therefore be used to search for proteins involved in a biological process of interest or to model the assembly of protein complexes by looking for clusters in PPI networks centered on the studied process. Second-order networks can be built and analyzed as well as first-order networks. APPINetwork provides information on PPIs in a large number of species or strains while other tools or databases are focused on a limited number of model species. Biologists working on lesser studied species or strains will therefore gain from using APPINetwork.

To remove protein data that may bias the clustering of a network, APPINetwork provides filtering options that are not offered by existing tools for building and analysis of PPI networks. Indeed, many proteins that are not specific to the biological process of interest are represented in second-order networks, while they interact with only one protein in the first-order network. If such proteins are not eliminated, the analysis tends to erroneously cluster proteins that have no biological relationship. Similarly, when looking for assembly intermediates of a protein complex, it can be useful to remove self-loops, because they are penalizing for the Jaccard index of dimeric proteins, which leads to assembly models where the monomeric proteins are assembled first.

APPINetwork removes inter species PPIs, which differentiates it from APID, PSICQUIC and mentha. This can be illustrated using the Elongator (Elp) complex of *S. cerevisiae* as an example. In the first-order networks built from the six subunits of the Elp complex with APID, PSICQUIC or mentha, one PPI is identified involving a human protein, namely the interaction between the protein ELP3 and the human histone H3.3. Notably, APPINetwork does not take interactions between a protein and another macromolecule into account. For example, according to the PSICQUIC network, the proteins ELP3 and IKI3 interact with the tRNA Glu UUC, but APPINetwork discounts these interactions.

APPINetwork merges all provided PPIs present in public and private PPI databases, so it builds a more complete network than other available tools. For example, by querying the BioGRID database, APPINetwork built a first-order PPI network of the ELP complex with more proteins related to the studied biological process than did APID, PSICQUIC or mentha. The additional identified proteins belong to transcription complexes TFIID, TFIIIB, SPT4-SPT5 and Facilitator of Chromatin Transcription (FACT) complex as well as the ribosome. None of these proteins are represented in the very small network obtained with STRING. Even when the STRING network is extended, it includes additional proteins with functions that do not seem to be coherent with those of the ELP complex proteins.

Using APPINetwork to integrate laboratory PPI data and PPI from public databases into the same network is particularly useful for analyzing the numerous PPIs identified by interatomic platforms. It will result in a more comprehensive network. An additional feature of APPINetwork is that the output contains information on the interactions of the network and the associated publication(s) describing them. This file is convenient for users because all known information on PPIs involving the proteins of interest is easily accessible.

An advantage of APPINetwork is that the user can build a PPI network with particular second-order PPIs, excluding more proteins that are unrelated to the biological process of interest. Moreover, as some proteins have a very many partners (several hundred), there is an option to filter out these partners which tend to strongly bias the clustering of the graph. The resulting clusters will thus be more relevant than when starting from a classical second-order network.

Finally, the first and second-order networks obtained with APPINetwork are provided in files that can be easily exported in other software like Cytoscape (*Shannon et al., 2003*) allowing them to be visualized and analyzed through other applications.

CONCLUSION

The APPINetwork package is a tool for PPI network building and analysis involving proteins from numerous biological processes in numerous species or strains. It offers users the choice of using public (experimental or predicted) PPI databases to build the PPI network and to add unpublished PPI data.

It has a user-friendly graphical interface allowing access to the different options for building a network suited to the type of analysis to be carried out. For example, a network built with genetic or predicted interactions, as well as unpublished interactions, could identify more PPIs involved in the studied biological process. A first-order network without self-loops could improve the likelihood of identifying assembly intermediates of a protein complex while a second-order network would identify sets of proteins involved in the same biological process. Other options of the interface allow to choose between the two types of analysis and modify their parameters.

APPINetwork provides the PPI network as a flat file containing the list of PPIs with various information about the interaction and the interacting proteins (PubMed IDs, experimental methods, all identifiers of involved proteins) that can be a very useful

resource for biologists. It also provides a text file containing all proteins of each cluster identified by TFit and additional files containing results of the hierarchical clustering modeling the assembly of a complex.

Finally, the APPINetwork package can be freely downloaded from the GitHub repository and comes with a user guide and examples that facilitate its use.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was funded by the University Evry-val-d'Essone (Fonds pour le rayonnement de la recherche 2020-Action 2). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
University Evry-val-d'Essone.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Simon Gosset performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Annie Glatigny conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Mélina Gallopin conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Zhou Yi performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Marion Salé performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Marie-Hélène Mucchielli-Giorgi conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The software is available at GitLab: https://forgemia.inra.fr/GNet/appinetwork.

REFERENCES

Alonso-López D, Campos-Laborie FJ, Gutiérrez MA, Lambourne L, Calderwood MA, Vidal M, De Las Rivas J. 2019. APID database: redefining protein–protein

interaction experimental evidences and binary interactomes. *Database* **2019**:baz005 DOI 10.1093/database/baz005.

Antezana E, Blondé W, Egaña M, Rutherford A, Stevens R, De Baets B, Mironov V, Kuiper M. 2009. BioGateway: a semantic systems biology tool for the life sciences. *BMC Bioinformatics* **10**:S11 DOI 10.1186/1471-2015-10-S10-S11.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. 2004. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **32**:D115–D119 DOI 10.1093/nar/gkh131.

Aranda B, Blankenburg H, Kerrien S, Brinkman FS, Ceol A, Chautard E, Dana JM, De Las Rivas J, Dumousseau M, Galeota E, Gaulton A, Goll J, Hancock RE, Isserlin R, Jimenez RC, Kerssemakers J, Khadake J, Lynn DJ, Michaut M, O'Kelly G, Ono K, Orchard S, Prieto C, Razick S, Rigina O, Salwinski L, Simonovic M, Velankar S, Winter A, Wu G, Bader GD, Cesareni G, Donaldson IM, Eisenberg D, Kleywegt GJ, Overington J, Ricard-Blum S, Tyers M, Albrecht M, Hermjakob H. 2011. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nature Methods* **29**:528–529 DOI 10.1038/nmeth.1637.

Bader GD, Cary MP, Sander C. 2006. Pathguide: a pathway resource list. *Nucleic Acids Research* **34**:D504–D506 DOI 10.1093/nar/gkj126.

Calderone A, Castagnoli L, Cesareni G. 2013. mentha: a resource for browsing integrated protein–interaction networks. *Nature Methods* **10**:690–691 DOI 10.1038/nmeth.2561.

Del Toro N, Anjali S, Ragueneau E, Meldal B, Combe C, Barrera E, Perfetto L, How K, Prashansa R, Shirodkar G, Lu O, Mészáros C, Watkins X, Sangya P, Licata L, Iannuccelli M, Pellegrini M, Martin MJ, Panni S, Duesbury M, Vallet SD, Rappsilber J, Ricard-Blum S, Cesareni G, Salwinski L, Orchard S, Porras P, Panneerselvam K, Henning H. 2022. The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Research* **50(D1)**:D648–D653 DOI 10.1093/nar/gkab1006.

Doncheva NT, Morris JH, Gorodkin J, Jensen LJ. 2019. Cytoscape StringApp: network analysis and visualization of proteomics data. *Journal of Proteome Research* **18(2)**:623–632 DOI 10.1021/acs.jproteome.8b00702.

Gambette P, Guénoche A. 2011. Bootstrap clustering for graph partitioning. *RAIRO-Operations Research* **45**:339–352 DOI 10.1051/ro/2012001.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**:R80 DOI 10.1186/gb-2004-5-10-r80.

Glatigny A, Gambette P, Bourand-Plantefol A, Dujardin G, Mucchielli-Giorgi MH. 2017. Development of an in silico method for the identification of subcomplexes involved in the biogenesis of multiprotein complexes in Saccharomyces cerevisiae. *BMC Systems Biology* **11**:67 DOI 10.1186/s12918-017-0442-0.

Glatigny A, Mathieu L, Herbert CJ, Dujardin G, Meunier B, Mucchielli-Giorgi MH. 2011. An in silico approach combined with in vivo experiments enables the identification of a new protein whose overexpression can compensate for specific respiratory defects in Saccharomyces cerevisiae. *BMC Systems Biology* **25**:173 DOI 10.1186/1752-0509-5-173.

Holmås S, Puig RR, Acencio ML, Mironov V, Kuiper M. 2019. The Cytoscape Bio-Gateway App: explorative network building from the BioGateway triple store. *Bioinformatics 9;* **36(6)**:1966–1967 DOI 10.1093/bioinformatics/btz835.

Humphreys IR, Pei J, Baek M, Krishnakumar A, Anishchenko I, Ovchinnikov S, Zhang J, Ness TJ, Banjade S, Bagde SR, Stancheva VG, Li XH, Liu K, Zheng Z, Barrero DJ, Roy U, Kuper J, Fernández IS, Szakal B, Branzei D, Rizo J, Kisker C, Greene EC, Biggins S, Keeney S, Miller EA, Fromme JC, Hendrickson TL, Cong Q, Baker D. 2021. Computed structures of core eukaryoticprotein complexes. *Science* **374**:6573 DOI 10.1126/science.abm4805.

Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**:299–314 DOI 10.2307/1390807.

Legeay M, Doncheva NT, Morris JH, Jensen LJ. 2020. Visualize omics data on networks with Omics Visualizer, a Cytoscape App. *F1000 Research* **9**:157 DOI 10.12688/f1000research.22280.2.

Martin A, Ochagavia ME, Rabasa LC, Miranda J, Fernandez-de Cossio J, Bringas R. 2010. BisoGenet: a new tool for gene network building, visualization and analysis. *Bioinformatics* **11**:91 DOI 10.1186/1471-2105-11-91.

Miryala SK, Anbarasu A, Ramaiah S. 2018. Discerning molecular interactions: a comprehensive review on biomolecular interaction databases and network analysis tools. *Gene* **642**:84–94 DOI 10.1016/j.gene.2017.11.028.

Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H. 2014. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research* **42(D1)**:D358–D363 DOI 10.1093/nar/gkt1115.

Orchard S, Hermjakob H. 2008. The HUPO proteomics standards initiative–easing communication and minimizing data loss in a changing world. *Brief Bioinformatics* **9**:166–173 DOI 10.1093/bib/bbm061.

Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, Dolma S, Coulombe-Huntington J, Chatr-Aryamontri A, Dolinski K, Tyers M. 2021. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science* **30(1)**:187–200 DOI 10.1002/pro.3978.

Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, Kolas N, O'Donnell L, Leung G, McAdam R, Zhang F, Dolma S, Willems A, Coulombe-Huntington

J, Chatr-Aryamontri A, Dolinski K, Tyers M. 2019. The BioGRID interaction database: 2019 update. *Nucleic Acids Research* **47**:D529–D541 DOI 10.1093/nar/gky1079.

Pan A, Lahiri C, Rajendiran A, Shanmugham B. 2016. Computational analysis of protein interaction networks for infectious diseases. *Brief Bioinformatics* **17**:517–526 DOI 10.1093/bib/bbv059.

Philipp O, Osiewacz HD, Koch I. 2016. Path2PPI: an R package to predict protein–protein interaction networks for a set of proteins. *Bioinformatics* **32**:1427–1429 DOI 10.1093/bioinformatics/btv765.

Porras P, Barrera E, Bridge A, Del-Toro N, Cesareni G, Duesbury M, Hermjakob H, Iannuccelli M, Jurisica I, Kotlyar M, Licata L, Lovering RC, Lynn DJ, Meldal B, Nanduri B, Paneerselvam K, Panni S, Pastrello C, Pellegrini M, Perfetto L, Rahimzadeh N, Ratan P, Ricard-Blum S, Salwinski L, Shirodkar G, Shrivastava A, Orchard S. 2020. Towards a unified open access dataset of molecular interactions. *Nature Communications* **11(1)**:6144 DOI 10.1038/s41467-020-19942-z.

Razick S, Magklaras G, Donaldson IM. 2008. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* **3**:405 DOI 10.1186/1471-2105-9-405.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**:2498–504 DOI 10.1101/gr.1239303.

Snider J, Kotlyar M, Saraon P, Yao Z, Jurisica I, Stagljar I. 2015. Fundamentals of protein interaction network mapping. *Molecular Systems Biology* **11**:848 DOI 10.15252/msb.20156351.

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering CV. 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47(D1)**:D607–D613 DOI 10.1093/nar/gky1131.

The UniProt Consortium. 2018. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **46**:2699 DOI 10.1093/nar/gky092.

Tran L, Hamp T, Rost B. 2018. ProfPPIdb: pairs of physical protein–protein interactions predicted for entire proteomes. *PLOS ONE* **13(7)**:e0199988 DOI 10.1371/journal.pone.0199988.

Wang L, Yang L, Peng Z, Lu D, Jin Y, McNutt M, Yin Y. 2015. cisPath: an R/Bioconductor package for cloud users for visualization and management of functional protein interaction networks. *BMC Systems Biology* **9(Suppl 1)**:S1 DOI 10.1186/1752-0509-9-S1-S1.

Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright G, Bader GD, Morris Q. 2010. The GeneMANIA prediction server: biological

network integration for gene prioritization and predicting gene function. *Nucleic Acids Research* **38**(**Web Server issue**):W214–W220 DOI 10.1093/nar/gkq537.

Wickham H, Hester J, Chang W. 2019. Devtools: tools to make developing R packages easier. *Available at https://cran.r-project.org/web/packages/devtools/index.html*.

3.3 . Une seconde version d'APPINetwork

Depuis la publication, plusieurs améliorations ont été apportées à APPINetwork par Alexandre Labesse, un étudiant de Master 1 de bioinformatique que j'ai co-encadré l'an dernier. Ces améliorations sont les suivantes :

3.3.1 . Construction automatique du fichier d'entrée

Dans la première version d'APPInetwork, l'utilisateur devait construire au préalable un fichier contenant trois identifiants de chacune de ses protéines intérêts (uniprotID, locus, nom de gène). Dans sa version 2, un bouton "Construct an input list" a été rajouté permettant à l'utilisateur de ne donner qu'un seul identifiant par protéine, les identifiants des différentes protéines pouvant être hétérogènes

3.3.2 . Téléchargement automatique des bases de données

Dans la première version d'APPInetwork, l'utilisateur devait préalablement télécharger des fichiers provenant de différentes bases de données de PPIs, avant de les interroger. Pour automatiser cela, nous voulions utiliser les API des différentes bases de données, mais les résultats générés étaient limités : par exemple l'API de BioGRID renvoie au maximum 10000 PPIs de S. cerevisaie sur les 800000 contenues dans la base de données. J'ai donc envisagé une boucle mais cela prenait trop de temps et une clef d'accès était demandée. On s'est donc dirigé vers des URL de téléchargement disponible pour chaque ressource. Dans la V2 d'APPINetwork, en cliquant sur le bouton "Format file" de l'interface, l'utilisateur peut désormais lancer un programme python qui va permettre de télécharger directement la base de données (Biogrid, Intact ou irefIndex) de l'espèce étudiée. Si la dernière version de la base de données n'existe pas dans le répertoire local de travail, alors une url est générée et le fichier est chargé et formaté sous R. À la fin, les bases de données formatées sont toutes enregistrées dans le répertoire Database du répertoire de travail. On peut suivre l'état d'avancement de tout le processus grâce l'affichage des messages dans la console R, et l'ensemble est stocké dans un fichier de suivi qui permet de retracer les différents paramètres qui ont été choisis par l'utilisateur..

3.3.3 . Visualisation dynamique de réseau

Dans la première version d'APPINetwork, le réseau était fourni sous forme d'un fichier .txt et d'un fichier .pdf donc non manipulable. Dans la V2, une visualisation dynamique du réseau est possible où l'utilisateur peut bouger les nœuds et changer les layouts. Pour arriver à cela, j'ai utilisé deux outils de visualisation : le package R tkplot qui permet d'effec-

tuer une visualisation rapide de réseau et Cytoscape, qui est un logiciel indépendant de R, pour lequel il existe un module R et qui permet des représentations plus avancées du réseau. J'ai ajouté un bouton "Network visualisation" au niveau de l'interface, qui redirige vers une fenêtre demandant à l'utilisateur le type de visualisation (tkplot ou cytoscape), ainsi que le fichier .txt contenant le réseau PPI. Une fois ces informations saisies, l'utilisateur peut cliquer sur le bouton "visualisation". Le fichier contenant le réseau PPI est alors chargé dans R puis transformé en graph afin d'être appelé par tkplot ou par cytoscape selon le choix du type de visualisation de l'utilisateur.

3.3.4 . Visualisation des clusters obtenus par Tfit

Dans le cas d'un réseau de PPIs de second ordre, analysé avec tfit, les protéines du réseau sont groupées en cluster. Dans la première version d'APPINetwork, le résultat de ce clustering était fourni sous la forme d'un fichier contenant une suite de lignes contenant les numéros des clusters et la liste des protéines composant ces clusters. Il est désormais visualisable sur le réseau. Pour faire cela, j'ai ajouté au réseau de second ordre deux colonnes contenant le numéro du cluster auquel chaque protéine appartient. Ainsi, pour une interaction entre une protéine A et une protéine B, j'ai ajouté le numéro du cluster de la protéine A en colonne 14 et celui de la protéine B en colonne 15. Lors de la visualisation, les numéros de cluster sont utilisés comme code couleur.

4 - Production des descripteurs des surfaces et des interfaces de des protéines en interaction

L'un des objectifs initiaux de ma thèse était de mettre en place une méthode de prédiction des PPI afin de pouvoir compléter l'interactome d'*A. thaliana* avec des PPI prédites, mais aussi de corriger le réseau en éliminant les faux positifs. C'est avec cet objectif en tête que j'ai cherché des descripteurs des surfaces et des interfaces protéine-protéine pouvant servir à la prédiction des interactions. J'ai ainsi calculé, pour un grand nombre de PPI, plusieurs types de cartes représentant la projection de différentes propriétés à la surface de chaque protéine en interaction ainsi que les énergies de liaison entre deux protéines en interaction. Ces cartes ont été obtenues à partir d'un outil et d'un pipeline que j'ai mis en place à partir du travail d'un ancien doctorant.

4.1 . SURFMAP : un logiciel pour mapper les propriétés physico-chimiques des protéines en 2 dimension

SURFMAP est un logiciel utilisable en lignes de commandes qui permet, à partir d'un fichier PDB de calculer un ensemble de propriétés physico-chimiques caractérisant la surface d'une protéine et de les projeter en deux dimensions. C'est un logiciel qui a été développé principalement par Hugo Schweke pendant sa thèse et auquel j'ai beaucoup contribué, ce qui m'a permis d'être 4eme auteur de l'article en annexe.

Ma contribution a été (1) de tester le logiciel sur un ensemble de protéines de *S. cerevisiae* de façon à identifier les bugs à partir des cartes produites, (2) de corriger les bugs et (3) de présenter le logiciel sous la forme d'un poster au congres JOBIM 2021. De plus certains scripts sont partagés entre SURFMAP et le pipeline IPOPS qui sera présenté dans la section suivante. J'ai donc indirectement participé au développement du code de SURFMAP en modifiant les scripts pour le pipeline IPOPS.

4.1.1 . Les propriétés physico-chimiques des surfaces protéiques prédéfinies dans le logiciel SURFMAP

SURFMAP permet à partir d'un fichier PDB de calculer un ensemble de propriétés physico-chimiques des surfaces protéiques :

- L'hydrophobicité des acides aminés à la surface. L'hydrophobicité d'un acide aminé est une valeur numérique reflétant à quel point

chaque acide aminé est hydrophobe ou hydrophile. Deux échelles d'hydrophobicité sont utilisées dans SURFMAP : (1) l'échelle Kyte-Doolitle où les valeurs associées à chaque acides-aminé ont été obtenues à partir d'un ensemble d'expérimentations faisant intervenir des solvants hydrophobes et hydrophiles [28]. Cette échelle est très utilisée pour détecter les segments transmembranaires ainsi que les régions hydrophobes à la surfaces des protéines. Comme décrit 1.2.2.3, la force hydrophobe est importante pour stabiliser la structure tertiaire des protéines et des complexes protéiques. (2) L'échelle Wimley-White qui mesure l'énergie libre associée au transfert d'une chaine d'un acide aminé d'un solvant eau à une membrane lipidique [76].

- La variance circulaire (CV) [39] permet de mesurer l'enfouissement d'un atome ou d'un acide aminé au sein d'une protéine. Elle est calculée en mesurant la densité atomique autour d'un atome ou d'un acide aminé. Cela donne une information importante sur la géométrie à la surface des protéines, notamment sur la convexité ou la concavité d'une région à la surface d'une protéine. La CV est comprise entre 0 et 1, une valeur faible indiquant que l'atome ou l'acide aminé est situé dans une région protubérante et une valeur faible indiquant qu'il est situé dans une cavité. La géométrie de la surface d'une protéine est bien évidemment importante pour l'interaction avec une autre protéine, même s'il est possible qu'à l'approche d'un partenaire potentiel, une protéine subisse des ré-arrangements structuraux.

- La stickiness [31] mesure la propension de chaque acide aminé à être impliqué dans une interface protéine-protéine.

- Le potentiel électrostatique à la surface est calculé grâce au logiciel APBS [23] qui est inclus dans SURFMAP. Cela permet d'associer à chaque atome d'une protéine sa charge électrostatique. Comme décrit 1.2.2.1, le potentiel électrostatique est très important pour la reconnaissance des partenaires protéiques.

4.1.2 . Projection des propriétés physico-chimique sur des cartes en deux dimension

Une fois que la propriété physico-chimique d'intérêt calculée ou qu'une propriété quelconque est encodée dans la colonne "b-factor" d'un fichier pdb, le logiciel MSMS [58] est utilisé dans SURFMAP pour générer un ensemble de points à 3Åtout autour de la surface de la protéine d'in-

térêt. Cet ensemble de points est appellé la "shell". La valeur de l'atome le plus proche est associé chaque particule de la "shell". Le fait qu'il y ait bien plus de particules que d'atomes permet d'obtenir des cartes en 2D sans trous. Ensuite, les coordonnées tridimensionnelles des particules sont transformées en coordonnées sphériques, avec comme référentiel le centre de masse de la protéine d'intérêt. Les coordonnées sphériques permettent de représenter la position d'une particule dans l'espace à l'aide de trois valeurs : la distance du centre de la particule au centre du repère r, un angle θ formé par l'axe z et le vecteur allant du centre du repère jusqu'au centre de la particule, un angle ϕ formé par le plan formé par les axes x et y et et le vecteur allant du centre du repère jusqu'au centre de la particule. Cependant une approximation est faite par SURFMAP et la composante r est considérée comme identique pour l'ensemble des particules. Cela a pour effet d'approximer la surface de la protéine par une sphère. Le Logiciel SURFMAP est donc approprié pour des protéines globulaires mais pas pour des protéines membranaires ou fibreuses dont la surface sera mal modélisée.

Ces points sont ensuite représentés sur un plan en deux dimensions représentant la surface de notre protéine grâce à une projection sinusoïdale qui a pour propriété de conserver les aires des surfaces (figure 4.2) Avec cette projection, les axes sont les suivants : $x = \phi \cdot \sin\theta$ et $y = 90 - \theta$. Les valeurs des abscisses vont de 0 à 360 et celles des ordonnées vont de 0 à 180. La carte ainsi obtenue est ensuite divisée en cellules d'une taille de 5 par 5. La valeur de chaque cellule correspond à la moyenne des valeurs associées à chaque particule. La valeur de chaque cellule peut ensuite être lissée en réalisant la moyenne des cellules adjacentes.

4.1.3 . Fichiers de sorties de SURFMAP

En sortie de SURFMAP, plusieurs fichiers sont générés : (1) les fichiers pdb où la colonne "b factor" a été remplacée par la propriété physico-chimique choisie, (2) les cartes des propriétés physico-chimiques au format png ou pdf et (3) le fichier .txt associé à chaque carte, où chaque ligne représente une cellule et où deux colonnes successives contiennent respectivement la valeur de la propriété physique associée à la cellule et la liste des acides aminés dont au moins un atome est présent dans la cellule.

4.2 . Le pipeline IPOPS

Le pipeline IPOPS est un pipeline réalisé en python et en R afin de générer des cartes de propension à l'interaction (IPOS) à partir des résultats de docking moléculaire. Initialement, les différents scripts de ce pipeline

avaient été écrits par Hugo Schweke durant sa thèse [59]. Mon travail sur le pipeline IPOPS a donc consisté à optimiser ces scripts, à en corriger les bugs, à y ajouter différentes fonctionnalités puis à les organiser en un pipeline permettant l'exécution successive des scripts. Suite à une collaboration avec Sjoerd De vries, ce travail sera ajouté au web server RPBS "Ressource Parisienne en BioInformatique Structurale (RPBS)" et valorisé comme tel dans un futur article.

4.2.1 . Docking moléculaire avec ATTRACT

Afin de construire la carte IPOPS d'une protéine *A*, le pipeline IPOPS demande en entrée les résultats du docking moléculaire de cette protéine *A*, considérée comme récepteur, avec un ensemble de protéines, considérées comme ligands. Au préalable, j'ai donc utilisé le logiciel de docking ATTRACT [74] pour calculer les énergies d'interaction entre deux protéines. ATTRACT est un logiciel de docking dit "gros-grains". En effet, il fait une approximation des structures des deux protéines afin de réduire l'espace conformationnel, ce qui permet de réduire considérablement le temps de calcul nécessaire pour réaliser le docking d'une paire de protéines. Au lieu de représenter l'ensemble des atomes des acides aminés, les concepteurs d'ATTRACT ont fait le choix de représenter les acides aminés par seulement 3 ou 4 pseudo-atomes. Deux pseudo-atomes servent à représenter les atomes participant à la chaîne principale, et un pseudo-atome ou deux pseudo-atomes permettent de représenter les atomes de la chaîne latérale, la glycine étant un cas particulier puisqu'elle ne possède pas de chaîne latérale et est donc représentée uniquement par les deux pseudo-atomes participant à la chaîne principale. Cette simplification permet de réduire la complexité des calculs tout en tenant compte des propriétés propres aux différents acides aminés, qu'elles soient conformationnelles ou physico-chimiques [74].

Pour un docking entre une protéine *A* que l'on appellera le récepteur et une protéine *B* que l'on appellera le ligand, ATTRACT fixe le récepteur au centre d'un repère et génère de manière homogène un ensemble de positions de départ pour le ligand autour du récepteur. Pour chacune des positions de départ du ligand, ATTRACT agit sur la position et l'orientation du ligand de façon à minimiser l'énergie d'interaction afin d'obtenir une position finale du ligand avec l'énergie la plus favorable à l'interaction. Mathématiquement cela revient à trouver le minimum local d'une fonction de score en agissant sur les 6 degrés de libertés rotationnels et translationnels.

Ainsi on obtient un point qui représente le centre de masse du ligand à sa position finale autour du récepteur, auquel est associé un minimum local de la fonction de score. L'opération est répétée un grand nombre

de fois pour de nouvelles positions de départ du ligand, et de nouveaux points associés à des minimums locaux sont trouvés (figure 4.2).

La fonction de score d'ATTRACT, qui sert à calculer l'énergie d'interaction, utilise le potentiel de Lennard-Jones, classiquement utilisé en docking pour représenter les forces de Van-der-Waals, et un terme de Coulomb pour représenter les forces électrostatiques pouvant entrer en jeu dans l'interaction entre le récepteur et le ligand.

Depuis l'écriture des scripts par Hugo Schweke, il y a eu plusieurs mises à jour d'ATTRACT permettant une accélération des temps de calculs. En premier lieu, la possibilité d'exécuter ATTRACT sur GPU, ainsi que la possibilité d'utiliser une grille. La grille est une représentation de l'espace autour du récepteur. Elle est divisée en voxels et au sein de chaque voxel, une estimation de l'énergie d'interaction est pré-calculée lorsque le voxel est occupé par le ligand. La grille permet ainsi de réduire drastiquement le temps de calcul nécessaire lors de l'étape de minimisation.

Les dockings générés par Hugo Schweke au cours de sa thèse [59] ont été générés avec la version CPU d'ATTRACT sans utilisation de la grille. Ce n'est pas le cas des dockings que j'ai générés avec la version GPU d'ATTRACT et avec la grille. Nous avons vérifié sur quelques exemples que les résultats trouvés avec la version GPU d'ATTRACT étaient similaire à ceux obtenus avec la version CPU figure (4.1).

Mais pourquoi avoir fait le choix d'ATTRACT parmi l'ensemble des méthodes de docking ? Lorsque le doctorant qui m'a précédé a développé la méthode IPOPS, il lui fallait une méthode de docking qui lui permette d'explorer l'ensemble de la surface d'un récepteur avec un ligand, dans un temps de calcul raisonnable, puisqu'il devait répéter ce calcul un grand nombre de fois. Il a alors réalisé l'inventaire des différentes méthodes disponibles. Ainsi, une méthode comme HADDOCK [84], qui est conçue pour être utilisée avec des contraintes qui guideront le processus de docking, ne permet pas d'explorer la surface protéique de manière homogène dans un temps raisonnable. D'autres méthodes comme ZDOCK [53] permettent l'exploration de millions de conformations différentes pour l'interaction entre deux protéines, mais seuls les meilleurs résultats sont disponibles à l'utilisateur, et leur répartition n'est pas homogène à la surface du récepteur. ATTRACT représentait donc le meilleur compromis pour une exploration exhaustive de l'énergie d'interaction à la surface d'une protéine dans un temps calcul raisonnable. Nous avons donc décidé de continuer à utiliser ATTRACT pour rester dans la continuité des résultats de la thèse de Hugo Schweke, et pouvoir utiliser les propriétés

intéressantes des cartes IPOPS qu'il a montrées durant sa thèse, telles que le recoupement des îlots favorables à l'interaction et des sites d'interactions fonctionnels des protéines. De plus, les arguments avancés par Hugo Schweke en faveur d'ATTRACT sont toujours valides aujourd'hui. En effet, pour les méthodes avec contraintes, la méthode de docking de référence est toujours HADDOCK et pour des méthodes permettant d'explorer des millions de conformations, c'est le serveur CLUSPRO [26] basé sur la méthode de docking PIPER qui ne permet de récupérer que les 10 meilleurs modèles ainsi que ZDOCK qui ne fournit également que les 10 meilleurs modèles.

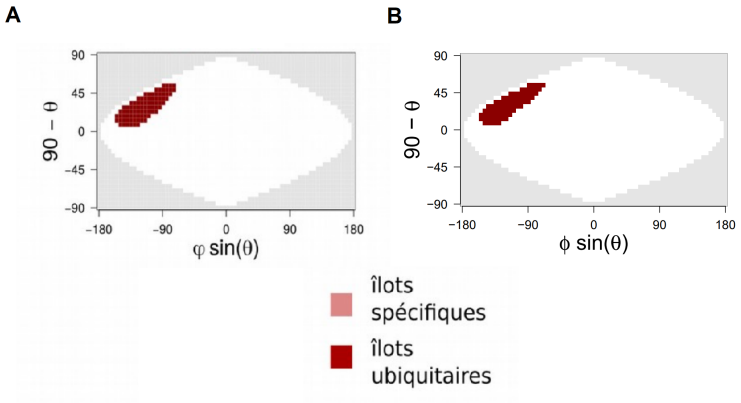


Figure 4.1 – **Comparaison des ilots obtenus sur une carte IPOPS pour la Structure 1GV3.** En A la carte IPOPS générée par Hugo au cours de sa thèse avec ATTRACT CPU, en B la carte générée avec ATTRACT GPU (figure reprise de [59]).

Avant de procéder au docking avec ATTRACT, les protéines sont préparées à l'aide de l'outil dockprep de la suite CHIMERA [29] qui permet l'élimination des molécules d'eau et des ions, la reconstruction des chaînes latérales incomplètes, l'ajout des atomes d'hydrogène et le traitement des ions non-standards. C'est une étape essentielle pour obtenir des résultats de docking fiables.

4.2.2 . Etape 1 du pipeline IPOPS : Calcul des cartes d'énergie

Une fois les calculs de docking réalisés, pour chaque docking récepteur-ligand :

Les coordonnées de chaque point représentant le centre masse du ligand à la position finale sont passées en coordonnées sphérique : Un angle θ, formé par l'axe z et le vecteur allant du centre de masse du récepteur au centre de masse du ligand, et un angle ϕ formé par le plan

formé par les axes x et y et et le vecteur allant du centre de masse du récepteur au centre de masse du ligand. L'énergie résultant du docking est associée au point et donc au couple d'angles (ϕ,θ) (figure 4.2).

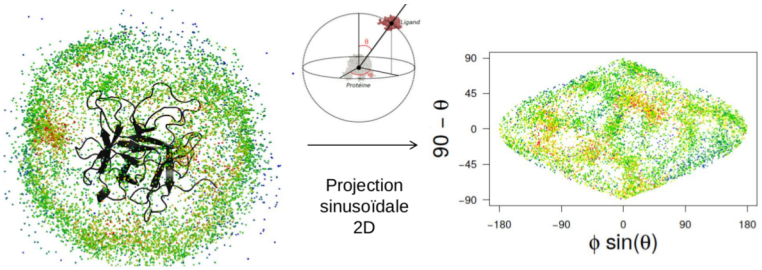


Figure 4.2 – **coordonnées sphériques et projection sinusoidale** (figure reprise de [59])

De la même manière que pour SURFMAP, Ces points sont projetés sur un plan en deux dimension grâce à une projection sinusoïdale qui a pour propriété de conserver les aires des surfaces (figure 4.2). La valeur de chaque cellule correspond à la moyenne des valeurs de l'ensemble des solutions de docking qui se trouve à l'intérieur d'une cellule. Un filtre est appliqué de façon à ne pas considérer les solutions de docking peu pertinente [59]. Ainsi, les points associés à une solution de docking avec une énergie de plus de 2.7 $Kcal.mol^{-1}$ sont ignorés pour le calcul de la moyenne. La carte ainsi obtenue est ensuite lissée en moyennant les valeurs de 8 cellules adjacentes. Lorsque la cellule n'a pas 8 cellules voisines de valeur supérieure à 0, la valeur de la cellule est passée à 0. Cela concerne les cellules présentes sur les bords de la cartes. Ces différentes étapes mènent à une carte en deux dimensions représentant les énergies de liaison entre le récepteur et le ligand à la surface du récepteur(figure 4.3).

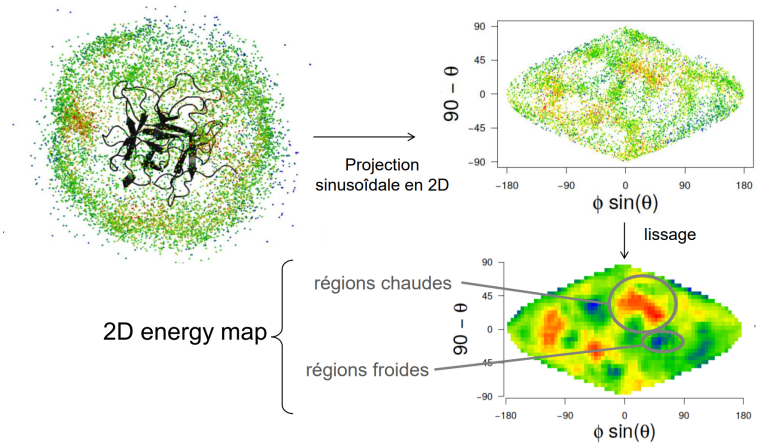


Figure 4.3 – **Lissage d'une carte d'énergie** (figure reprise de [59])

4.2.3 . Etape 2 du pipeline IPOPS : discrétisation des cartes d'énergies

Les valeurs des cartes d'énergies sont discrétisées en 5 classes de même amplitude : (5) très favorable, (4) favorable, (3) neutre, (2) défavorable, (1) très défavorable à l'interaction entre le ligand et le récepteur. Les gammes d'énergies sont calculées selon la formule suivante :

$$[max(E) - Cl_{range} \times (i);\ max(E) - Cl_{range} \times (i-1)] \qquad (1)$$

avec

$$Cl_{range} = |max(E) - min(E)|/5$$

où *i* est la classe d'énergie, allant de 1 à 5, *max(E)* et *min(E)* respectivement le score d'énergie le plus défavorable et le plus favorable à l'interaction de la carte d'énergie.

Les valeurs des cellules de la carte sont remplacées par le numéro de la classe à laquelle sa valeur appartient. On obtient ainsi une carte de classes d'énergie dont les valeurs des cellules vont de 1 à 5. (figure 4.4)

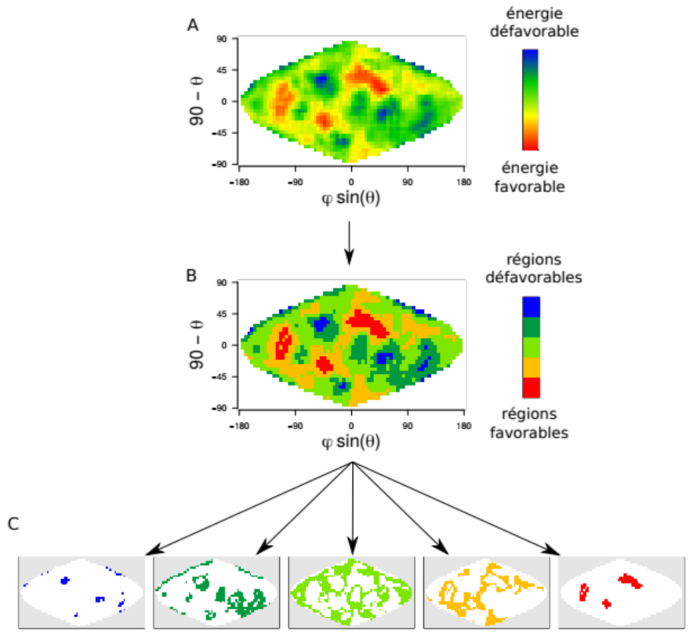


Figure 4.4 – **Discrétisation des cartes d'énergie en carte de classes d'énergie et séparation en 5 cartes binaires** (figure reprise de [59])

4.2.4 . Etape 3 du Pipeline IPOPS : Calcul des cartes IPOPS

Les cartes IPOPS d'un récepteur sont calculées à partir de l'ensemble des cartes de classes d'énergies de ce récepteur docké contre différents ligands. Une carte IPOPS sera réalisée pour chaque classe d'énergie. Comme il y a 5 classes d'énergies, il y aura donc 5 cartes IPOPS par récepteur.

Dans un premier temps, chaque carte de classes d'énergie est divisée en 5 cartes binaires contenant des 1 si la cellule contient *i* et 0 sinon.
Dans un second temps, pour une classe d'énergie donnée, les cartes binaires de classe *i* du récepteur avec ses différents ligands sont sommées. Chaque cellule de cette carte contient donc le nombre de ligands dont l'énergie de liaison avec ce récepteur à la position donnée par la cellule est de classe d'énergie *i* (figure 4.5).

Dans un troisième temps, les cartes sont normalisées de façon à toujours être sur une échelle de 0 à 100 quelque soit le nombre de ligands dockés contre le récepteur (figure 4.5).

Enfin, une étape d'élimination du bruit de fond est réalisée de façon à ne conserver que des zones et non pas des points isolés (pour plus de détails voir [59]).*i*. La figure 4.6 en donne une illustration.

Les cartes ainsi obtenue sont appelées cartes IPOPS. Pour chaque ré-

cepteur, 5 cartes sont produites, une par classe d'énergie. Pour la catégorie, la plus favorable, l'intensité des pixels (la valeur dans la cellule) représente la propension d'une zone (d'un îlot) à la surface du récepteur à être favorable à l'interaction pour toutes les protéines du set de ligands avec lesquelles les dockings ont été réalisés. A l'inverse, pour la classe d'énergie la plus défavorable, l'intensité d'un pixel représente la propension d'une zone (d'un ilôt) de la surface du récepteur à être défavorable à l'interaction.

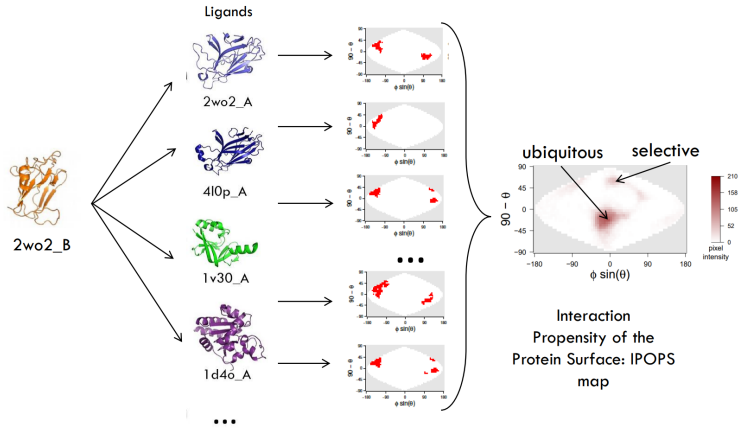


Figure 4.5 – **Construction d'une carte IPOPS** (figure reprise de [59])

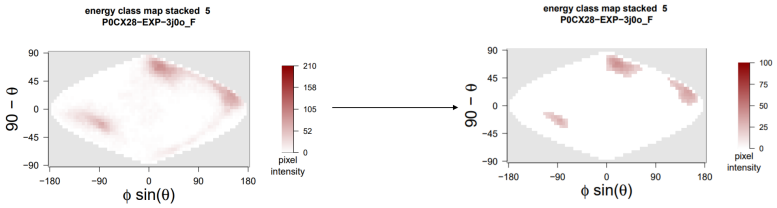


Figure 4.6 – **Normalisation et élimination du bruit de fond d'une carte IPOPS** (figure reprise de [59])

4.2.5 . Cartes rose/rouge

Au cours de sa thèse, Hugo Schweke a identifié que sur les cartes IPOPS de la classe d'énergie la plus favorable aux interactions (carte rouge), deux populations d'îlots peuvent être identifiées [59]. Des îlots dont l'intensité est élevée (îlots rouges) et des îlots dont l'intensité est plus faible (îlots roses). Les îlots rouges représentent des zones de la surface du récepteur très favorable à l'interaction avec un grand nombre de ligands, les îlots roses représentent des zones de la surface du récepteur

très favorable à l'interaction avec un petit nombre de ligands (figure 4.6)

Ces îlots roses représenteraient donc des zones très favorables à l'interaction mais plus spécifiques. Pour la prédiction des PPIs, ces îlots roses pourraient ainsi permettre de différencier une interaction non spécifique et non fonctionnelle pour la cellule, d'une interaction fonctionnelle. En effet ces zones roses pourraient représenter des sites d'interactions spécifiques à la surface des récepteurs et parmi le petit nombre de protéine pouvant interagir sur ce site, se trouveraient des protéines pouvant former des interactions fonctionnelles. Dans l'environnement cellulaire qui est très encombré et dans lequel toutes les protéines sont en contact les unes avec les autres, les interactions fonctionnelles seraient ainsi favorisées par leur plus grande spécificité au niveau de ces îlots roses.

4.2.6 . Etape 4 du Pipeline IPOPS : Extraction des îlots

Le pipeline IPOPS fournit des fichiers listant les coordonnées des pixels associés aux différents îlots présents sur les cartes IPOPS de la classe d'énergie la plus favorable à l'interaction (carte rouge).

Il fournit aussi des fichiers listant les coordonnées des pixels associées aux différents îlots présent sur les cartes IPOPS rose/rouge.

4.2.7 . Les amélioration apportées au pipeline IPOPS

Outre la correction de bugs, mon travail sur le pipeline IPOPS a contribué à certains ajouts notables tel que :

- L'optimisation du pipeline permettant la parallélisation des étapes qui prenaient le plus de temps. Le choix du nombre de coeurs autorisé pour la parallélisation est en option.
- L'ajout de nombreuses options telles que la projection Mollweide. Comme la projection sinusoïdale, la projection Mollweide permet de conserver l'aire des surfaces. Elle est très utilisée dans les projections du globe terrestre dans son entièreté et présente une forme ovale. La distorsion des différentes formes sur la carte n'est pas la même qu'avec la projection sinusoïdale. Les formes sont moins déformées sur les bords droit et gauche de la carte, cependant la projection sinusoïdale conserve mieux les formes au niveau du méridien central.
- La possibilité de générer les cartes au format pdf à toutes les étapes du pipeline.
- La possibilité de pointer la position du centre de masse d'un ou de plusieurs acides aminés sur la carte en 2 dimensions représentant la surface du récepteur.
- La correspondance entre les cellules et les acides aminés du récepteur. Au tout début du pipeline, juste avant le passage en coordon-

nées sphériques, chaque point correspondant au centre de masse du ligand est associé à l'acide aminé le plus proche sur le récepteur. Cela permet, lors des étapes suivantes, de savoir quelles sont les différents acides aminés présents dans chacune des cellules représentant la surface du récepteur, ou encore de savoir quels acides aminés sont impliqués dans les îlots roses et rouges.

Dans sa thèse, Hugo Scwheke a montré que les sites d'interactions des partenaires protéique formant un complexe de structure connue se trouvent dans les îlots des cartes IPOPS de classe 5, c'est à dire les zones très favorables à l'interaction. Il a ainsi montré que les cartes IPOPS pouvaient servir à identifier les sites d'interaction entre des partenaires connus pour former des PPI. Elles doivent donc contenir des informations utiles à la prédiction des paires de protéines en interaction.

4.3 . Résultats

4.3.1 . Résultats de Docking chez *Saccharomyces cerevisiae*

Lors de sa thèse Hugo Schweke avait réalisé le docking croisé avec ATTRACT de l'ensemble des protéines globulaires de *Saccharomyces cerevisiae*, dont la structure était disponible dans la PDB. Une procédure de docking croisé consiste à réaliser le docking de toutes les paires de protéines formées à partir des protéines d'un jeu de données. Ce jeu de structure provient de la base de données interactome3D [42]. Le site interactome3D propose au téléchargement un set représentatif de structures protéiques pour différents organismes modèles comme *Saccharomyces cerevisiae* ou encore *Arabidopsis thaliana*. Dans ces sets représentatifs, la structure d'un complexe est soit (1) la structure expérimentale la plus complète du complexe parmi celles fournies par la PDB, soit (2) un modèle généré par homologie. Lorsque les structures étaient incomplètes incomplètes et que la partie manquante se trouvaient au niveau d'une boucle, ces boucles ont été modélisées. Les structures expérimentales incomplètes ont été utilisées pour le docking croisé lorsque la boucle modélisée faisait moins de 20 acides aminés. Certains modèles ont été générés à partir de structures homologues incomplètes. Dans ce cas, seuls ceux contenant des boucles modélisées de moins de 5 acides aminés ont été utilisés pour le docking croisé. Ces données ont été générées à la fin de la thèse de Hugo Schweke avec ATTRACT en utilisant les paramètres décrits dans la table 4.1.

Cependant on ne savait pas exactement à quel point le docking croisé réalisé était complet. J'ai donc fait un inventaire de ces données de docking. Il y avait initialement 12 to de données qu'il a fallu trier. Certains fichiers de docking étaient complètement vides, d'autres n'avaient pas été traités par ATTRACT pour l'élimination des positions redondantes. Un

grand nombre de fichiers de docking existaient en plusieurs exemplaires. De plus, les fichiers étaient compressés dans des archives qui dans certains cas étaient elle même dans des archives.

Pour faire le tri dans les fichiers, ma stratégie a été de construire une matrice résumant les docking réalisés. Au sein de cette matrice les lignes représentent les protéines considérées comme récepteurs lors du docking par ATTRACT et les colonnes représentent les protéines considérées comme les ligands. Pour chaque résultat de docking identifié dans un des répertoires, la case correspondante à la paire récepteur-ligand est cochée, et les fichiers de résultats sont déplacés dans nouveau répertoire. Pour 27 226 fichiers de dockings identifiés, j'ai relancé les dernières étapes d'ATTRACT qui consiste à éliminer la redondance des positions au sein des fichiers de dockings, conformément au protocole d'Hugo Schwek décrit dans la table 4.1e.

Une fois ce travail de nettoyage et d'inventaire réalisé, j'ai obtenu une matrice de 471 récepteurs par 471 ligands. Cette matrice est asymétrique. En effet, c'est le centre de masse du récepteur qui est l'origine du repère utilisé pour identifier la position du ligand à la surface du récepteur. Ainsi une carte où une protéine A est considérée comme récepteur par ATTRACT et une protéine B est considéré comme ligand n'est pas équivalente à une carte ou cette même protéine B est le récepteur et la protéine A le ligand. La matrice que j'ai obtenue est remplie à 63% (Fig :4.7). Il manque donc malgré tout un certains nombre de dockings pour que le docking croisé soit complet. Parmi les 471 protéines, 4 protéines n'ont jamais été dockées en tant que récepteurs.

Etape de minimisation	**Seuil (Å^2)**	V_{max}
1	1500	50
2	500	60
3	150	60
4	80	100
5	80	100

Table 4.1 – **Options utilisées pour la procédure de Docking de Hugo Schweke.** La V_{max} correspond au nombre maximal de pas de minimisation autorisé pendant l'étape de minimisation d'énergie. Le seuil correspond au carré de la distance seuil maximale (en Å^2) autorisée entre deux pseudo-atomes pour l'établissement des listes de pseudo-atomes mises à jour au début de chaque étape de minimisation d'énergie considérés pour la minimisation.

Si on observe la distribution du nombre de ligands dockés par récepteurs en excluant les 4 protéines qui n'ont jamais servi de récepteur, on

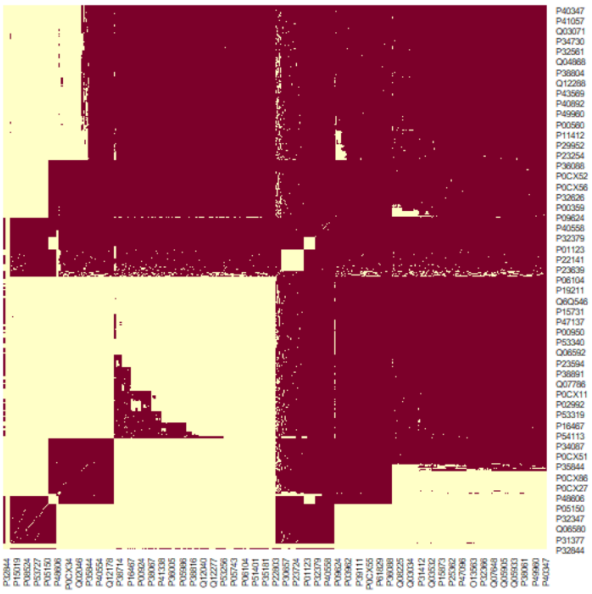


Figure 4.7 – **Matrice présentant le nombre de dockings récupérés.** Les lignes représentent les protéines considérées comme récepteurs lors de la procédure de docking par ATTRACT. Les colonnes représentent les protéines considérées comme les ligands lors de la procédure de docking par ATTRACT.

constate que les récepteurs pour lesquelles le moins de ligands ont été dockés ont au minimum été dockés avec 90 ligands (figure 4.8). Durant sa thèse, Hugo avait étudié la variabilité des cartes IPOPS en fonction du nombre de ligands utilisés pour les générer. Son études s'était surtout intéressée au cartes IPOPS de catégories 5 (très favorables à l'interaction). Il avait estimé que 20 ligands suffisaient pour obtenir des cartes robustes pour l'identification des acides aminés participant au site d'interaction.

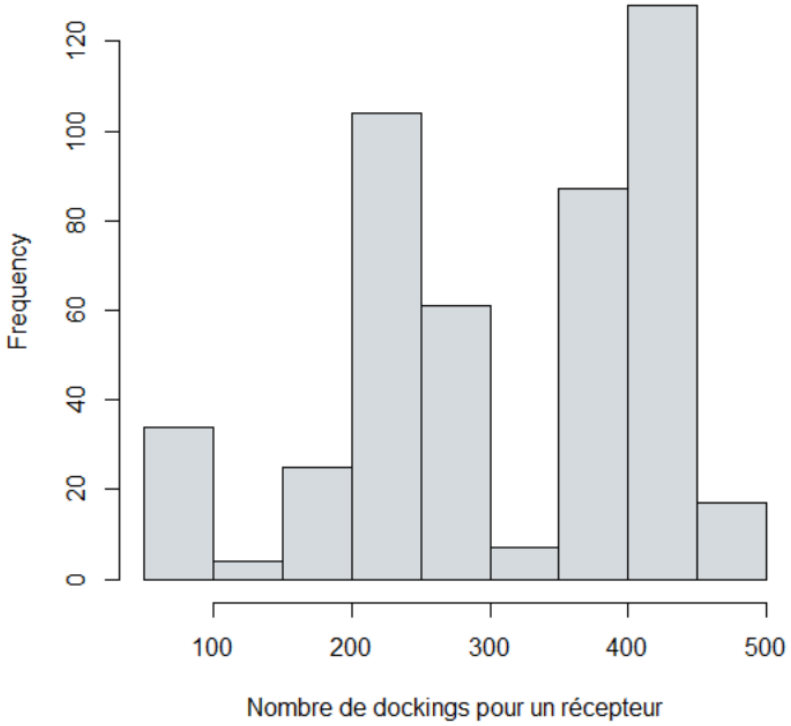


Figure 4.8 – **Distribution du nombre de ligands par récepteur.** Les récepteurs ont été dockés au minimum avec 90 ligands. En moyenne, ils ont été dockés avec 316 ligands.

4.3.2 . Générations des cartes IPOPS chez *S. cerevisiae*

Une fois l'inventaire des fichiers de docking réalisé chez *S. cerevisiae*. J'ai exécuté le pipeline IPOPS pour les 467 (471-4) récepteurs de la matrice. Cela m'a permis de générer 2335 cartes IPOPS (5 cartes par récepteur, une par catégorie d'énergie).

Options	Valeurs
Nombre de classes	5
Taille de la grille	5
Type de projection	Sinusoïdale
Normalisation	Utilisation d'une liste de valeurs

Table 4.2 – **Options utilisées pour le pipeline IPOPS.** Le nombre de ligands dockés par récepteur n'étant pas le même, une liste de valeurs correspondant aux nombre total de ligands dockés pour chacun des récepteur a été utilisée afin d'avoir une normalisation correcte.

A l'aide du logiciel SURFMAP, j'ai également généré pour ces 467 récepteurs, les cartes des propriétés physico-chimiques suivantes : la stickiness, le potentiel électrostatique, l'hydrophobicité avec l'échelle Kyte-Doolitle, la variance circulaire par atome et par acide aminé.

4.3.3 . Données de docking générées chez *A. thaliana*

Pour réaliser une procédure de docking croisé chez *A. thaliana*, il m'a fallu tout d'abord rassembler un ensemble de structures. Dans cet objectif j'ai rassemblé les structures expérimentales présentes dans le jeu de données "représentatif" de la base de données Interactome3D ([42]).

J'ai ensuite sélectionné les structures des protéines globulaires sur la base de leur localisation subcellulaire lorsqu'elle était renseignée dans la base de données uniprot, notamment en filtrant les protéines sur la base de mots clés tel que "membrane". Cela donne un total de 213 structures d'*A. thaliana* globulaires.

Les structures ont ensuite été préparées pour le docking en suivant le même protocole que Hugo Schweke, c'est à dire en utilisant le protocole de DockPrep de la suite CHIMERA [29].

Le docking croisé des 213 protéines d'*A. thaliana* a été réalisé avec une procédure de docking différente de celle de Hugo Schweke (voir table suivante) car c'est la version ATTRACT GPU qui a été utilisée pour réduire le temps de calcul.

Etape de minimisation	Seuil (Å^2)	V_{max}
1	Grille	1000

Table 4.3 – **Options utilisées pour la procédure de Docking ATTRACT GPU.** L'utilisation de la grille rend inutile la réalisation de plusieurs étapes de minimisation comme réalisées dans le protocole d'Hugo Schweke. Avec la grille précalculée au début de la procédure de docking, la valeur du seuil pour l'établissement des listes de pseudo-atomes mises à jour au début de chaque étape de minimisation est considérée comme infinie et l'entièreté de la surface de la protéine est considérée. Une seule étape de minimisation est donc réalisée avec un nombre de pas (V_{max}) plus important.

4.4 . Les raisons et les limites d'une prédiction de PPI basée sur les cartes d'énergies, les cartes de propension à l'interaction et les cartes de propriétés des surfaces

Pour prédire les PPI, nous avons pensé utiliser les cartes d'énergies de docking, couplées aux cartes de propension à l'interaction (IPOPS) et de propriétés de surfaces produites avec Surfmap. L'utilisation des cartes de propension à l'interaction pour prédire l'interaction entre deux partenaires est motivée par le recouvrement des îlots favorables à l'interaction avec les sites d'interaction des protéines [59]. En effet, s'il y a un lien entre les îlots présents sur les cartes et les surfaces par l'intermédiaire desquelles les partenaires protéiques interagissent les un avec les autres, on peut penser que l'information portée par les cartes IPOPS devrait être prise en compte pour prédire les PPIs. Pour interagir, deux protéines doivent avoir à leur surface deux zones favorables à l'interaction qui aient des propriétés de surface compatibles et dont l'énergie de liaison soit basse. Mais l'environnement de ces zones favorables est aussi important car si à proximité de ces zones favorables, se situe des zones défavorables à l'interaction qui vont être répulsives ou provoquer des clashs stériques, l'interaction ne pourra avoir lieu. Nous avons donc pensé faire apprendre à un réseau de convolution, les cartes de propension à l'interaction et de propriétés de surface ainsi que les valeurs d'énergies de docking correspondant aux différentes zones, pour des paires de protéines aléatoires et pour des partenaires protéiques déterminés expérimentalement.

Mais l'approche envisagée présente des limites :

- Tout d'abord, il est nécessaire de connaître les structures tridimensionnelles des deux protéines dont on souhaite prédire l'interaction. Le nombre de structures protéiques déterminées expérimentalement étant limité, il restreint le nombre de PPIs qu'il est possible

d'étudier. Pour contourner ce problème, il est possible d'enrichir le jeu de données par des modèles structuraux. Ces modèles sont déterminés soit (1) par homologie, c'est à dire que le modèle est calqué sur les structures expérimentales des protéines homologues qui ont été conservées au cours de l'évolution, soit (2) prédits par Alphafold2 [21] et mis à disposition dans la base de données AlphafoldDB (pour plus de détails, voir chapitre 5).

- Malgré l'utilisation d'un logiciel docking gros grains (ATTRACT) et donc moins coûteux en temps de calcul pour réaliser le docking, la génération des cartes IPOPS nécessite un temps de calcul important. L'utilisation de la version GPU d'ATTRACT a permis de réduire ce temps de calcul mais celui-ci reste cependant une limite importante de cette approche.

4.5 . Les méthodes utilisant AlphaFold2 pour la prédiction de PPI permettent de surmonter ces limites

Fin 2021, de nouvelles méthodes de prédiction de PPI utilisant AlphaFold2 ont été publiées [5, 20]. Elles permettent de prédire une PPI à partir de la séquence des deux protéines sans avoir besoin de leurs structures et avec un temps de calcul inférieur au temps de génération de cartes IPOPS. De plus, ces méthodes prédisent la structure du complexe, qui peut être utilisée a posteriori, pour étudier plus en détail les caractéristiques des surfaces et des interfaces des protéines en interaction. Par ailleurs, les performances de ces méthodes (précision, Taux de Faux positif (FPR) et Recall) surpassent largement les précédentes. En effet, l'état de l'art des méthodes de prédiction de PPI décrit dans l'article [15] publié en 2021 avant la publication des méthodes utilisant AlphaFold2, montre que les meilleures performances atteignaient 0.279 de recall pour un FPR de 0.01 (performances obtenues par la méthode EVcomplex2). On verra dans le chapitre 5 qu'elles seront bien meilleures par la suite.

La méthode présentée dans [20] utilise une stratégie couplant deux méthodes de prédiction : elle calcule une probabilité d'interaction pour tout le jeu de données avec la méthode Rosettafold [1] qui est très rapide, puis utilise Alphafold2 pour prédire uniquement les complexes des paires les plus probables. Cette stratégie permet ainsi l'étude d'un grand jeu de données très déséquilibré (avec 1000 fois plus de paires aléatoires que de paires en interaction, comme c'est le cas si on considère toutes les interactions possibles entre les protéines d'une espèce telle que la levure *S. cerevisiae*).

L'autre méthode [5], plus ciblée sur la prédiction de la structure des complexes que sur la prédiction des paires de protéines en interaction,

permet d'atteindre 0.51 de Recall pour 0.01 de FPR, tout en obtenant des structures de complexes hétérodimériques de très bonne qualité (62.7 % de modèles prédits avec un DockQ $\geq$ 0.23).

Pour tous les avantages présentés précédemment par les méthodes de prédiction utilisant AlphaFold2 qui lèvent les limites d'une prédiction basée sur les cartes IPOPS et les cartes d'énergie de docking, tout en permettant d'obtenir d'excellents résultats de prédiction, nous avons décidé d'utiliser AlphaFold-Multimer [11] (la version d'AlphaFold2 dédiée à la prédiction de complexes protéiques) pour prédire des PPI, puis d'utiliser les structures prédites des complexes pour étudier les propriétés de leurs interfaces en utilisant les cartes produites avec SURFMAP et IPOPS. Ainsi, si des propriétés propres aux PPI qui n'ont pas été correctement prédites émergent, elles pourront être utilisées a posteriori afin d'améliorer les performances de prédiction.

L'objectif de ma thèse étant d'enrichir un interactome par des interactions prédites, la méthode de prédiction doit être adaptée pour un vaste jeu de données. Nous avons donc choisi d'appliquer la méthode présentée dans [20] qui a été développée dans cet objectif.

5 - Utiliser Alphafold2 pour prédire des PPI

5.1 . L'arrivée de Alphafold2, véritable révolution pour la biochimie structurale

Le domaine de la prédiction des structures des protéines a été marqué par une grande révolution technologique aux cours des dernières années. En 2018, a lieu la 13^{eme} édition d'un concours sur le sujet de la prédiction des structures des protéines, le concours CASP (Critical assessment of structure prediction) [27]. L'objectif du concours est de permettre une évaluation objective et comparative des méthodes de prédiction de structures protéiques développés par des chercheurs du monde entier. Pour cela, des structures résolues expérimentalement par la communauté scientifique sont rassemblées sans être publiées, de façon à ce que les équipes participantes soit confrontées à des structures de protéines qui ne sont pas encore présentes dans les bases de données de structures protéique. Les participants utilisent alors leurs méthodes de prédiction de façon à prédire les structures tridimensionnelle de ces protéines. Les structures prédites sont ensuite comparées aux structures expérimentales de façon à évaluer les performances des méthodes mises en place par les participants. Cette 13^{me} édition du concours est marqué par la participation de l'entreprise Deepmind. En effet, Alphafold, la méthode mise en place par Deepmind lors de cette édition, montre d'excellentes performances [61], les meilleurs du concours, notamment en ce qui concerne la tache la plus difficile à réaliser, le "template free modelling" (FM) qui consiste à prédire la structure de protéines qui n'ont pas d'homologue avec une structure connue. C'est un défi plus grand car les prédictions sont faites sans partir d'un template. En 2020, Deepmind réitère sa participation au concours CASP pour la 14^{eme} édition avec leur méthode, Alphafold2. Si les performances d'Alphafold était déjà excellentes par rapport à ce qui se faisait en 2018, Alphafold2 présente des performances de prédiction exceptionnelles, auxquelles personne ne s'attendait, rivalisant même avec des méthodes de résolutions de structure expérimentale telle que la cristallographie par rayons X [22]. En 2021, deepmind met à disposition de la communauté scientifique une base de données qui contient aujourd'hui plus de 200 millions de structures protéiques prédites.

Véritable révolution dans le domaine de la biochimie structurale, l'impacte d'Alphafold2 s'est rapidement étendu à d'autre domaine de la protéomique, notamment le domaine de la prédiction de PPI. En 2021, un article publié par Humphrey et al [20] montre qu'il est possible d'utiliser Alphafold2 pour prédire des PPI avec une précision élevé. Cela nous a

amené à revoir notre stratégie de prédiction et à utiliser Alphafold2 pour notre problème de prédiction de PPI.

5.2 . Alphafold2

Alphafold2 réalise la prédiction de la structure tridimensionnelle d'une protéine à partir de sa séquence d'acides aminés, grâce à une méthode d'apprentissage profond basée sur un système d'attention multicouche. Les méthodes basées sur l'attention multicouche sont très utilisées dans tout ce qui à trait au traitement du langage et ont été popularisées suite à la publication du célèbre papier "Attention is All you Need" [72] qui pose les bases de l'architecture particulière utilisée dans le réseau de neurones Alphafold2 : le transformer.

L'attention est un mécanisme qui permet à des méthodes d'apprentissage supervisées, tels que les réseaux de neurones, de se "concentrer" sur certaines parties d'une séquence d'entrée lors du traitement de l'information, comme les humains prêtent attention à des parties spécifiques d'un ensemble d'éléments lors de la compréhension ou de la réalisation d'un problème. On parle de mécanisme d'attention multicouche lorsque la méthode de prédiction peut "se concentrer" sur plusieurs partie de la séquence d'entrée simultanément pour réaliser sa tache de prédiction. Ainsi avec des mécanismes d'attention, le réseau n'apprend pas seulement comment réaliser une prédiction, il apprend aussi quelle partie de la séquence d'entrée est pertinente pour la prédiction.

Alphafold2 réalise sa prédiction à partir de la séquence d'acides aminés d'une protéine d'intérêt. A partir de cette séquence, Alphafold2 construit une matrice d'alignement multiple (MSA) et une représentation "par paire". L'information contenue dans ses deux représentations sera affinée dans le premier module principal d'Alphafold2, l'evoformer. Le module structurel qui est le deuxième module principal d'Alphafold2 va réaliser la prédiction de la structure tridimensionnelle à partir des matrices affinées par l'evoformer (figure 5.1).

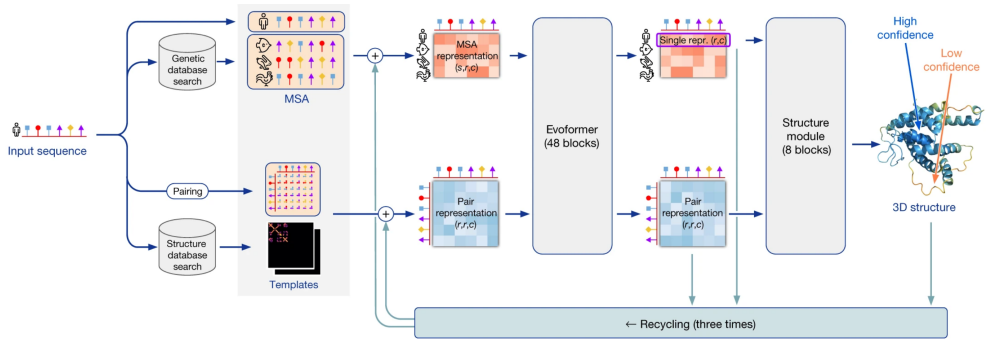


Figure 5.1 – **Schéma du fonctionnement de Alphafold2** (figure reprise de [22]).

5.2.1 . La matrice d'alignement multiple (MSA)

La matrice d'alignement multiple (MSA) est la matrice obtenue par l'alignement de plusieurs séquences protéiques de manière à maximiser la similarité ou la conservation des acides aminé afin de refléter une évolution commune.

Dans le cas d'Alphafold2, la construction de la MSA repose sur l'alignement de la séquence de la protéine dont on veut connaître la structure 3D avec des orthologues chez d'autres espèces. Un orthologue est un gène qui provient d'un ancêtre commun à différentes espèces et qui conserve une fonction similaire chez toutes ces espèces. L'intérêt d'inclure des orthologues dans la MSA est de mettre en lumière les acides aminés importants pour la conservation de la structure de la protéine d'intérêt. En effet, deux acides aminés en contact dans la structure 3D de la protéine et essentiels à son maintien auront tendance au cours de l'évolution soit à être conservés à l'identique, soit à subir tous les deux une mutation de façon à ce que le contact entre les deux acides aminés soit toujours possible et que la structure et la fonction de la protéine puisse elle aussi être conservée (voir figure 5.2). Ainsi, la MSA doit contenir un nombre de séquences suffisantes, réparties sur l'histoire évolutive de la protéine d'intérêt, afin de capturer suffisamment d'informations sur les mutations et les conservations qui se sont produites au fil du temps pour mettre en valeur les acides aminés en contact dans sa structure 3D. Cependant, si une MSA contient trop de séquences, elle ne permettra pas non plus de mettre en valeur les paires d'acides aminés ayant coévolués car l'information évolutive est trop bruitée. Il semblerait qu'une MSA d'une taille de 30 séquences soit suffisante pour une prédiction correcte de la structure 3D d'un monomère [22].

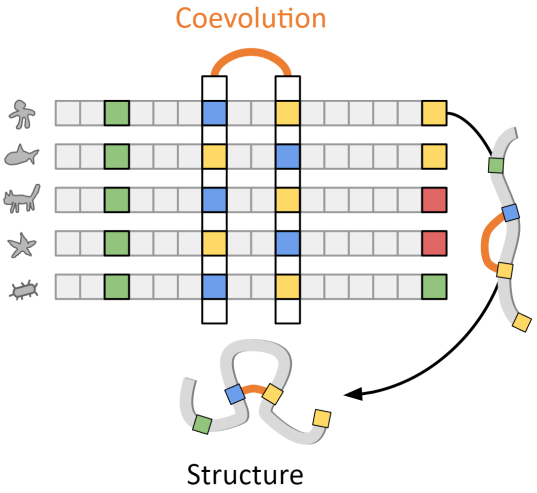


Figure 5.2 – **Schéma de l'utilisation de la coévolution pour la prédiction de contact dans une MSA.** La première ligne correspond à la séquence en acides aminés d'une protéine dont on veut prédire la structure, les autres lignes correspondent aux séquences en acides aminés d'orthologues de la séquence d'intérêt (figure reprise de [40]).

Un autre point qui peut poser problème est la différenciation des orthologues et des paralogues. Les paralogues sont des gènes qui ont évolués par duplication au sein d'un même génome et peuvent avoir des fonctions divergentes ou additionnelles. L'information évolutive risque donc d'être bruité par l'ajout de paralogues dans la MSA.

5.2.2 . Les "templates" structuraux

Une recherche dans des bases de données est réalisée de façon à trouver des protéines avec une séquence proche de la protéine d'intérêt et dont une structure a déjà été résolue expérimentalement. Ces structures vont servir de modèle, appelé "template". Les templates structuraux peuvent aider à améliorer la précision des prédictions en fournissant une base sur laquelle Alphafold2 peut s'appuyer. Néanmoins les templates ne sont pas nécessaire à la prédiction, en témoigne les très bonnes performances d'Alphafold2 dans la catégorie free modelling lors du concours CASP14.

5.2.3 . La représentation "par paire"

La représentation "par paire" est initialement construite à partir de la MSA. Lorsque la prédiction est réalisée avec template, l'information obtenue grâce aux templates y est ajoutée. C'est une représentation de la relation spatiale entre les différents acides aminés qui composent la structure

de la protéine d'intérêt.

5.2.4 . Embedding

L'"embedding" est une technique couramment utilisée en apprentissage automatique. Dans le cas d'Alphafold2, il permet de transformer la MSA, qui est un alignement de séquence d'acides aminés, en une représentation numérique qui capture l'information pertinente pour la prédiction de la structure de la protéine d'intérêt. Ce dont le réseau se sert au sein de l'evoformer n'est donc pas directement la MSA mais une représentation numérique de la MSA.

L'embedding est aussi ce qui permet d'obtenir la représentation "par pair" à partir des templates sructuraux.

5.2.5 . L'Evoformer

L'Evoformer est la structure qui permet de capturer l'information contenue dans la représentation de la MSA et dans la représentation "par pair". C'est un processus itératif qui permet de raffiner les informations contenues dans ces deux représentations notamment grâce à des mécanismes d'attention. Une des innovations majeures est la communication permise entre la MSA et la représentation "par paire". L'information contenue dans la MSA est utilisée pour raffiner l'information structurale contenue dans la représentation "par paire", de même l'information structurale permet de raffiner l'information évolutive contenue dans la MSA (figure 5.1).

5.2.6 . Le module structural

Le module structural est le module qui va prédire la structure tridimensionnelle de la protéine d'intérêt. La prédiction est réalisée à partir de la première ligne de la MSA, qui correspond à la représentation de la séquence d'intérêt, et à partir de la représentation "par paire". (figure 5.1)

Au sein du module structural, la protéine d'intérêt est considérée comme un "nuage de résidus" : Chaque acide aminé est considéré comme un triangle rigide qui représente les atomes N-Cα-C qui est déplacé de façon à former la chaîne principale de la protéine d'intérêt. Les contraintes liées aux liaisons peptidiques peuvent parfois être rompues lors du déplacement des résidus d'acides aminés.

Une étape de relaxation sera cependant nécessaire après la prédiction pour que les différentes contraintes géométriques de la liaison peptidique soit respectées au niveau de la structure prédite. Pour cela, le champ de force Amber est appliqué à la structure prédite.

Hormis la représentation en nuage de résidus, une des grandes innovation de ce module est la mise en place d'un processus d'attention spécialement conçu pour les structures 3D. Ce processus, appelé "Invariant

Point Attention" (IPA), a la propriété très intéressante d'être invariant aux rotations et aux translations. Cela permet au modèle de ne pas considérer les structures quasiment identiques ayant pour seule différence une rotation ou une translation.

5.2.7 . Recyclage

A la fin d'un passage complet dans Alphafold2, la première ligne de la représentation de la MSA, la représentation par paire ainsi que la structure 3D obtenue après le module structural sont réinjectées au début du réseau pour un nouveau passage, de façon à affiner la structure prédite obtenue. Par défaut, le nombre de recyclage d'Alphafold2 est de 3 pour les monomères (figure 5.1).

5.2.8 . Mesures de qualités

Pour chacune de ses prédictions, Alphafold2 fournit plusieurs mesures de qualité de la structure qu'il a prédite. En temps normal, lorsqu'on veut mesurer la qualité d'une structure prédite, on la compare à la structure résolue expérimentalement. Dans le cas d'Alphafold2, des réseaux de neurones ont été entraînés de façon à prédire des mesures de qualités pour toutes les structures prédites.

- **Le predicted Template Modeling score (pTM) :** lorsque la structure expérimentale est connue, le TM score est utilisé pour mesurer la similarité entre un modèle prédit et la structure résolue expérimentalement. Pour cela les deux structures sont alignées l'une avec l'autre avec un algorithme d'alignement de structure, puis le TM score est calculé avec la formule 5.1. Un TM-score autour de 0,20 signifie généralement que les deux structures alignées correspondent à une paire de protéines choisie aléatoirement. Un TM-score supérieur à 0,5 suggère généralement une similitude structurale significative (au-delà de ce qui serait attendu par hasard [81]). Alphafold2 a été entraîné de façon à prédire une valeur de TM score : le pTM [22].

$$\text{TM-score} = \max\left(\frac{1}{L_{\text{target}}}\sum_{i=1}^{L_{\text{aligned}}}\frac{1}{1+\left(\frac{d_i}{d_0(L_{\text{target}})}\right)^2}\right) \tag{5.1}$$

Figure 5.3 – **Illustration de la superposition d'un modèle et d'une structure de référence pour calculer le TM score.** (figure reprise de bioinfo-fr.net).

où L_{target} est la longueur de la séquence de la structure expérimentale, L_{aligned} est le nombre d'acides aminés alignés, d_i est la distance entre les acides aminés alignés dans les structures comparées, enfin $d_0(L_{\text{target}})$ est une valeur de normalisation qui dépend de la longueur de la séquence de la protéine expérimentale, calculé suivant l'équation 5.2. La normalisation par d_0 permet d'avoir un score de similarité indépendant de la longueur de la séquence en acides aminés.

$$d_0(L_{\text{target}}) = 1.24 \sqrt[3]{L_{\text{target}} - 15} - 1.8 \tag{5.2}$$

- **Le predicted Local Distance Difference Test (pLDDT) :** lorsque la structure expérimentale est connue, le Local Distance Difference Test (LDDT) est utilisé pour mesurer la différence entre la structure expérimentale et le modèle prédit de manière locale. Il y a une valeur de LDDT calculée par acide aminé de la protéine prédite. Pour calculer le LDDT d'un acide aminé, les distances entre toutes les paires de Cα des acides aminés l'environnant sont évaluées dans le modèle et dans la structure expérimentale. Pour chaque distance, l'écart entre le modèle et la structure est calculé. Les distances sont considérées

comme identiques entre le modèle et la structure expérimentale si l'écart est inférieur à un certain seuil dit de "tolérance". Le pourcentage de distances environnantes qui sont identiques est calculé pour des seuils de tolérance de 0.5 Å, 1 Å, 2 Å et 4 Å. Le LDDT final est la moyenne de ces 4 pourcentages, voir figure 5.4. Alphafold2 a été entraîné de façon à prédire une valeur de LDDT, le pLDDT, pour la structure prédite [22]. Le pLDDT est une mesure locale de la qualité de la prédiction qui ne tient pas compte du placement et de l'organisation des domaines protéiques les un par aux autres au sein de la structure 3d 5.4. Elle peut varier entre 0 et 100. La confiance que l'on peut accorder à la qualité locale de la prédiction est la suivante : si le pLDDT est supérieur à 90, cette confiance est très élevée. Entre 90 et 70 cette confiance est élevée ; il faut cependant se méfier de la prédiction des chaîne latérales des acides aminés. Entre 70 et 50 cette confiance est faible. Un pLDDT inférieur à 50 est un excellent prédicteur de région intrinsèquement désorganisées. La confiance à accorder à la prédiction de la structure est alors très faible.

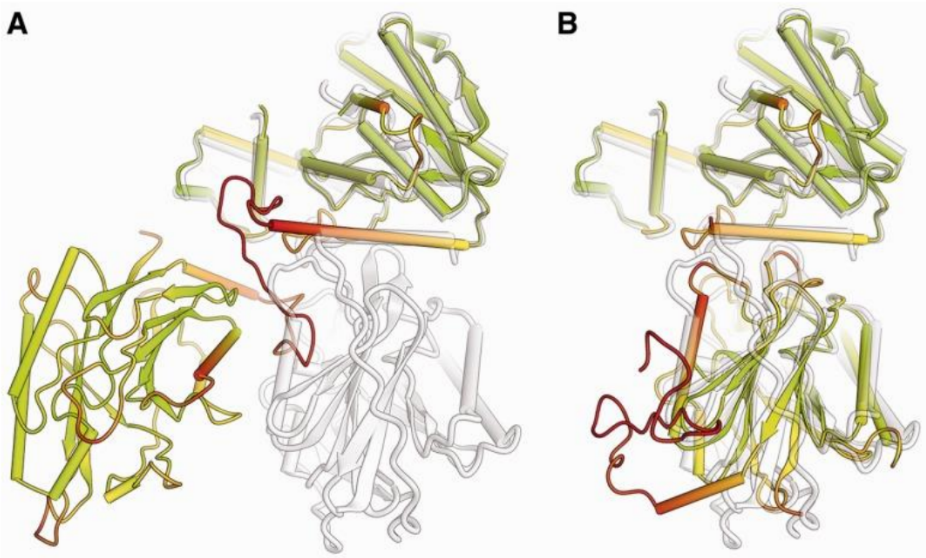


Figure 5.4 – **Illustration du LDDT.** La structure de référence est en transparent, le modèle est coloré le long de la chaîne protéique de façon à représenter la valeur du LDDT, en vert un LDDT élevé et en rouge un LDDT faible. En A la structure est superposée à partir des acides aminés d'un seul des deux domaines. On remarque que la valeur du LDDT au sein de chacun des deux domaines est élevée malgré le fait qu'ils ne se superposent pas, illustrant la mesure locale de la qualité du modèle. En B les domaines du modèle et de la structure de référence sont superposés indépendemment l'un de l'autre (la chaîne protéique a été coupée sur la boucle liant les deux domaines). La superposition des deux domaines est correcte là ou la valeur du LDDT est élevée. (figure reprise de [35].

- **La matrice Predicted Alignment Error (PAE) :** lorsque la structure expérimentale est connue, la matrice d'erreur d'alignement (PAE) est calculée en alignant le modèle avec la structure expérimentale avec comment point d'ancrage un résidu y, puis en calculant l'écart des distances pour chaque paire de résidus x du modèle et de la structure expérimentale. Puis l'alignement est réalisé sur le résidu y+1 et l'écart des distances pour chaque paire de résidus x est calculé de nouveau. On obtient ainsi une matrice de taille (n,n) avec n le nombre d'acide aminé de notre structure, voir figure 5.5. Dans chaque cellule de la matrice, la valeur correspond à un écart de distance en Å pour le résidu x entre la structure prédite et expérimentale lorsque les deux structures sont alignées avec comme point d'ancrage le résidu y. Alphafold2 a été entrainé pour prédire cette matrice afin d'obtenir la matrice de PAE. C'est une métrique qui est complémentaire du pLDDT. En effet si le pLDDT permet d'évaluer la

qualité locale de la prédiction, la matrice de PAE permet d'évaluer la qualité globale de la prédiction, et notamment si l'organisation des domaines les un par rapport aux autres a été correctement prédite, voir figure 5.5.

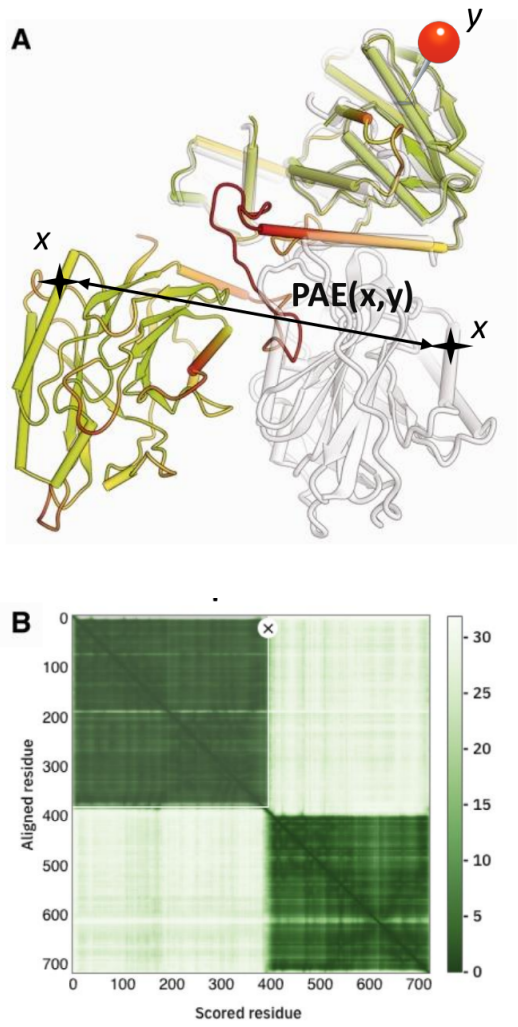


Figure 5.5 – **Illustration du calcul de la matrice de PAE A.** La structure de référence est en transparence et le modèle est coloré selon la valeur du LDDT (en vert un LDDT élevé et en rouge un LDDT faible). Le modèle et la structure de référence ont été superposés avec comme point d'ancrage l'acide aminé y, représenté par l'épingle rouge. Le PAE(x,y) est la distance entre l'acide aminé x de la structure de référence et du modèle, alignés sur y. On remarque que malgré un LDDT élevé, l'erreur d'alignement pour les acides aminés du domaine 2 est grande lorsque les structures sont superposées avec comme point d'ancrage, un acide aminé du premier domaine. Ceci illustre une erreur ou différence d'organisation des domaines. **B.** Un exemple de matrice de PAE(x,y) avec x en abscisses et y en ordonnées. On remarque ici quatre blocs, deux en vert foncé (PAE faibles) et deux en vert clair (PAE élevés) qui indiquent que les domaines sont bien prédits mais que l'organisation des domaines est erronée.

5.3 . Une première méthode de prédiction de PPI utilisant AlphaFold2

En décembre 2021, dans un article publié par Humphrey et colaborateurs [20], Alphafold2 est détourné de son utilisation première afin de prédire des complexes protéiques et des PPI chez *S. cerevisiae* . Pour prédire des complexes avec Alphafold2, les séquences des deux protéines dont on veut prédire l'interaction sont collées l'une avec l'autre en ajoutant entre elles un linker. Ce linker correspond à une succession de Glycine, ce qui permet d'avoir une boucle très flexible qui ne gênera pas la prédiction de la structure du complexe. Cependant, la construction de la MSA telle que mise en place par Alphafold2 n'est pas adaptée pour prédire des complexes. En effet la prédiction de complexe nécessite la construction d'une matrice de "paired MSA" (pMSA) (figure 5.6).

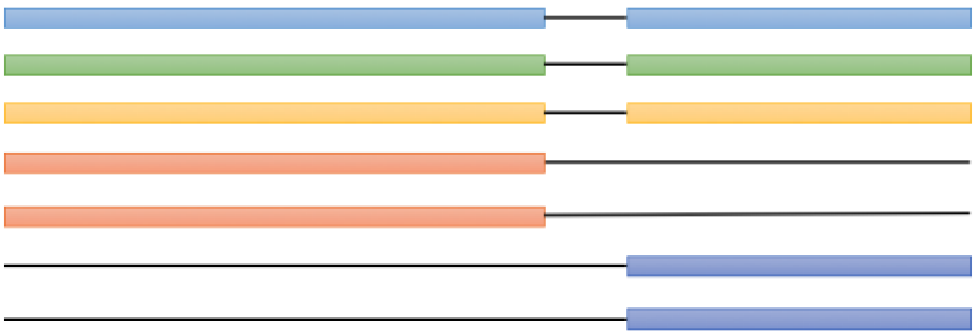


Figure 5.6 – **Schéma de l'obtention d'une pMSA par concaténation de deux MSA.** Chaque couleur indique une espèce différente. (figure reprise de [20]).

5.3.1 . La matrice de paired Multiple Sequence Alignment (pMSA)

Pour construire un complexe entre une protéine A et une protéine B, une recherche des protéines orthologues est effectuée pour la protéine A d'un coté et pour l'a protéine B de l'autre. L'objectif étant d'identifier des orthologues de la protéine A et des orthologues de la protéine B. L'alignement est réalisé pour la protéine A et la protéine B séparément. Enfin les deux MSA sont concaténées l'une avec l'autre de façon à ce que les séquences des orthologues de la protéine A et de la protéine B chez une même espèce se retrouve sur la même ligne (figure 5.6).

De la même manière que lors de la prédiction de la structure d'un monomère, la pMSA contient de l'information sur les résidus en contact au sein de la séquence de la protéine A, de la protéine B et aussi entre les deux séquences. Par exemple, si au cours de l'évolution, une mutation

sur un acide aminé de la protéine A est souvent accompagné par une mutation d'un acide aminé de la protéine B, cela pourrait signifier que pour conserver une interaction entre la protéine A et B une mutation compensatrice doit avoir lieu, et donc que ces deux résidus sont en contact lors de la formation du complexe.

Des problèmes supplémentaires peuvent arriver lors de la construction d'une pMSA, par exemple comment gérer les cas où un orthologue a été trouvé pour la protéine A chez une espèce mais pas pour la protéine B, ou alors comment gérer les cas où il y a plusieurs orthologue potentiels chez une espèce pour les protéines A et B. Plusieurs stratégies sont disponible pour ce genre de cas, mais dans l'article d'Humphrey et al, dans ces deux cas la concaténation n'est pas réalisé pour les séquences concernées. Elles sont ajoutées à la fin de la pMSA avec des gaps à la place de la protéine manquante car cela pourrait améliorer la prédiction des contacts intra-chaine.

Une fois ces matrices construites, elles sont utilisées en entrée de la méthode Alphafold2 pour prédire la structure des complexes.

5.3.2 . Calcul de la probabilité de contact

Alphafold2 est une méthode de prédiction de structure. Néanmoins parmi les matrices qu'il produit, il y en a une en particulier, qui est intéressante pour prédire des PPI : la matrice de distances. C'est une matrice qui contient les probabilités pour chaque paire de résidus d'avoir une distance entre leurs Cβ comprise dans un intervalle de distance donné. Il y a 64 intervalles répartis de 2 à 22 Å [22]. Pour un couple de protéine A et B dont on veut prédire l'interaction, il y a donc 64 matrices de dimension *(n+m,n+m)*, *n* et *m* étant les longueurs respectives des protéine A et B. Les probabilités de contact entre deux acides aminés sont calculées en sommant les probabilités, ce qui revient à calculer la probabilité que leurs Cβ se trouvent à moins de 12 Å (figure 5.7). La probabilité de contact la plus élevée est utilisée pour prédire l'interaction entre A et B [20].

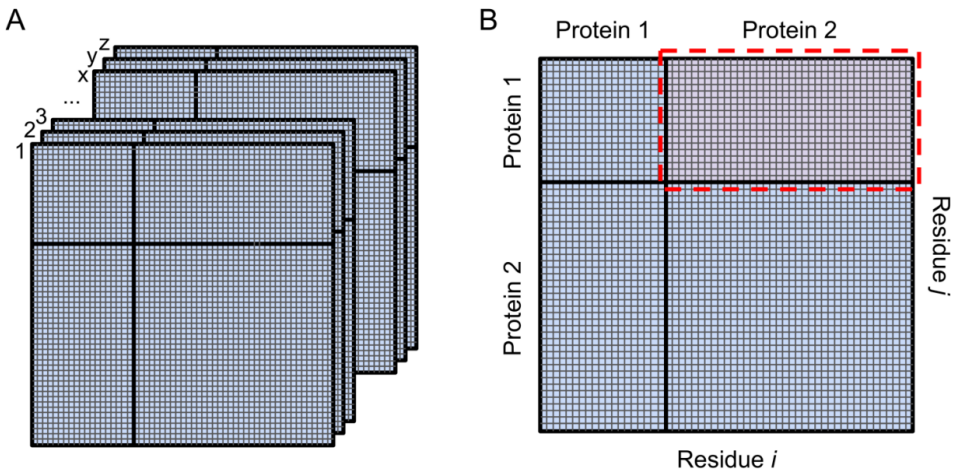


Figure 5.7 – **Schéma des matrices utilisées dans le calcul des probabilités de contactes.** (figure reprise de [20]).

5.3.3 . Performance

Dans [20], l'utilisation d'Alphafold2 a été couplée à une autre méthode : RoseTTAFold. RoseTTAFold est une méthode très inspirée de Alphafold2. Cependant, si elle est moins précise, RoseTTAFold a l'avantage d'être 100 fois plus rapide. Ainsi pour permettre une exploration à l'échelle d'un interactome, RoseTTAFold est utilisé en amont pour obtenir une probabilité de contact selon un procédé très similaire à celui décrit dans le paragraphe précédent. Si cette probabilité de contact est supérieure à un certain seuil alors seulement la prédiction de cette interaction est effectuée avec Alphafold2. Sinon on considère que les deux protéines ne peuvent interagir. Ainsi, RoseTTAFold agit ici comme un filtre permettant de ne présenter à Alphafold2 que les paires de protéines les plus pertinentes.

Les performances de la méthode RoseTTAFold+Alphafold2 ont été évaluées sur un jeu "Gold Standard" de 768 paires de protéines dont l'interaction est bien documentée. Pour ce qui est du jeu de données négatif (non interactions), 768 000 paires de protéines ont été générées aléatoirement en vérifiant bien que les paires générées ne soit pas référencées dans le jeu "Gold Standard", mais aussi dans différentes base de données de PPI (notamment Biogrid et STRING). Cette différence de taille entre les sets positifs et négatifs a pour objectif de reproduire la différence entre le nombre de non-interactions et le nombre de PPI que l'on pourrait retrouver lorsque l'on considère l'ensemble des paires de protéines possible d'un protéome (voir partie 1.2.4).

Cette combinaison de RoseTTAFold et d'Alphafold2 permet d'atteindre une précision de 0.95 pour un recall de 0.29. Ainsi, les PPI qui sont prédites

sont prédites avec une précision élevée, mais seulement 29% des PPI du "gold standard" sont correctement prédites, les autres étant prédites en tant que non-interactions.

Bien que conçu initialement pour prédire la structure tridimensionnelle des protéines monomériques, AlphaFold2 montre une capacité remarquable à prédire des structuctures de complexes. Les performances limitées de la méthode d'Hymphrey et al. peuvent être causées par l'utilisation de RoseTTAFold, qui permet certes, d'explorer une proportion plus grande de paires de protéines mais qui est moins performante qu'Alphafold2. Je me suis donc demandé si cette étape de présélection des paires les plus probables par RoseTTAFold était vraiment nécessaire et si on ne pouvait pas mettre en place une méthode ne passant pas par RosettaFold, qui soit aussi performante et suffisamment rapide pour permettre d'explorer un très grand nombre de PPI.

5.4 . Utilisation de Colabfold et Alphafold-Multimer pour prédire des PPI

J'ai donc décidé de mettre en place une méthode de prédiction de PPI similaire mais utilisant uniquement Alphafold-multimer [11], une version de Alphafold2 spécifiquement entraînée pour la prédiction de complexes protéiques, que l'on espère donc être plus adaptée pour la prédiction des PPI à l'échelle de l'interactome . Un problème subsiste cependant, la longueur des temps de calculs pour la construction des pMSA et la prédiction de la structure. Afin de répondre à ce problème, c'est une version un peu particulière de Alphafold-multimer que j'ai utilisée, Alphafold-multimer implémenté par Colabfold [40].

5.4.1 . Alphafold-Multimer

Il existe plusieurs différence entre Alphafold-Multimer et Alphafold2. Ces différences sont :

- le jeu d'entraînement qui a été utilisé : structures monomériques pour Alphafold2 et structures de multimériques pour Alphafold-Multimer.

- la construction des MSA. La prédiction de complexes avec Alphafold-Multimer nécessite la construction de pMSA et non de simple MSA.

- la fonction de perte. La fonction de perte d'une méthode de prédiction est une fonction qui, lors de l'entrainement, sert à mesurer l'écart entre la prédiction et la structure de référence. Les

paramètres de la méthode sont modifiés lors de l'entraînement de façon à optimiser cette fonction de perte et réduire l'écart entre la prédiction et la structure de référence. Alphafold-Multimer utilise une version modifiée de la fonction de perte utilisé dans Alphafold2. Les changements réalisés cherchent à améliorer les prédiction dans le cas de la prédiction de complexe et impliquent notamment : (1) un terme pénalisant la trop proche proximité des différentes chaînes protéiques afin d'éviter des superpositions et (2) un plafonnement plus haut des écarts distances pour les paires inter-chaînes par rapport au paires intra-chaînes.

- une métrique pour mesurer la qualité des interfaces prédites, **interface predicted Template Modeling score (ipTM)**. Lorsque la structure expérimentale est connue, l'iTM peut être utilisé pour mesurer la qualité de la prédiction entre des chaînes protéiques d'un modèle par rapport à une structure de référence. Pour calculer l'iTM, pour une chaine A, la structure de référence et le modèle sont alignés sur un acide aminé i de la chaine A, puis un TM score est calculé en se servant uniquement des résidus qui n'appartiennent pas à la chaine A. L'alignement est ensuite réalisé sur le résidu suivant et le TM score est calculé à nouveau en se servant des résidus qui n'appartiennent pas à la chaine du résidu i. Cette opération est réalisée sur l'ensemble des résidus du complexe. L'iTM sélectionné est l'iTM maximum sur l'ensemble des alignements réalisés. Dans le cadre d'Alphafold-Multimer, cette mesure est prédite par le réseau et est fournie comme une métrique pour mesurer la qualité des prédiction de l'interface entre les différentes sous-unités d'un complexe.

5.4.2 . Colabfold

Colabfold est une approche qui a pour but de rendre plus accessible Alphafold2 et Alphafold-Multimer notamment en proposant une méthode de construction des pMSA bien plus rapide, sans grande perte au niveau de la qualité de la prédiction [40]. Colabfold est disponible en utilisant les ressources de Google-colab, des ressources mises à disposition de la communauté scientifiques par google, ou en version locale, ce qui nécessite de posséder les ressources de calculs nécessaires pour réaliser les prédictions. Nous avons choisi d'utiliser la version locale puisque nous disposons de nos propres ressources de calculs sur le cluster de l'IPS2 (NVIDIA RTX A5000, possédant 24 Go de mémoire graphique). Cette capacité élevée en mémoire graphique permet de prédire la structure de complexes allant jusqu'à 2000 acides aminés au total.

5.4.2.1 . La construction des pMSA avec Colabfold

Pour réduire le temps de calcul d'Alphafold-Multimer, colabfold propose une méthode de génération des pMSA différentes. Premièrement en agissant sur les bases de données de séquences qui vont être parcourues pour rechercher les orthologues. Ainsi, une base de données ColabfoldDB a été créée à partir des bases de données BFD et MGnify utilisé classiquement par Alphafold-Multimer. BFD est une de base qui comprend 2,2 milliards de séquences protéiques, regroupées dans des clusters en fonction de leur similarité. Il y en a 64 million en tout. MGnify est une base de données qui comprend 300 millions de séquences protéiques. Pour réduire la taille de ces 2 bases de données, les séquences de la base MGnify ont été assignées au 64 millions de familles de protéines de BFD lorsque cela était possible, sinon de nouvelles familles de protéines ont été crées. Puis les 10 séquences les plus diverses de chaque famille ont été gardées. Cela permet de réduire drastiquement la taille des bases de données à parcourir. Ainsi la taille des 2 bases cumulées est réduite de 2,5 milliards de séquences à 513 millions. En utilisant le même principe, ces bases de données ont été étendues avec des séquence provenant de différentes bases de données de séquences d'Eukaryotes, d'ADN de virus et de bactériophages.

Cette méthode permet d'assurer une certaine diversité au niveau des séquences qui vont être utilisées pour construire les pMSA tout en réduisant la taille des bases de données et donc en accélérant les temps de calcul.

Deuxièmement, l'algorithme utilisé pour générer les alignements n'est pas le même. Originalement les algorithmes utilisés pour réaliser les alignements sont HMMer et HHblits, des méthodes très sensibles pour détecter des homologues. Colabfold permet l'utilisation de MMseqs2, une méthode d'alignement moins sensible mais beaucoup plus rapide. De plus, Colabfold met à disposition un serveur publique sur lequel il est possible de construire les pMSA à partir de ColabfoldDB. En utilisant le serveur, la construction de la pMSA prend genéralement une dizaine de minutes contre plusieurs heures en local sur notre cluster.

Les alignements sont réalisés séparément pour chaque chaine protéique, et des MSA sont générées. Si on a les séquences des protéines A et B dont on veut prédire le complexe, pour générer la pMSA les séquences sont pairées si et seulement si pour une espèce donnée, un homologue a été trouvé pour les deux séquences. Si plusieurs séquences ont été trouvées chez la même espèce pour la chaine A ou la chaine B, alors seule la séquence avec la meilleure e-value et qui couvre au moins 50% de la séquence d'intérêt est pairée afin qu'il y ait uniquement une paire d'homologues de la chaine A et B par espèce dans la pMSA.

La pMSA ainsi que les deux MSA non pairées sont fournies à Alphafold-

Multimer dans le cas de la prédiction de complexe avec Colabfold. Si la pMSA fournit des informations sur les contacts entre les résidus appartenant à des chaines protéique différentes, les MSA non pairées fournissent de précieuses informations sur les contacts entres les résidus au sein de la même chaine protéique.

5.4.2.2 . Performance de Colabfold dans le cadre de la prédiction de structures

Dans l'ensemble, les performances de Colabfold à la fois sur la prédiction de structure de monomère et de complexes sont très proches de celles d'Alphafold2 et Alphafold-Multimer. Cependant le temps d'exécution nécessaire pour obtenir une prédiction à partir des séquences de protéines d'intérêts est bien plus rapide. Dans l'article, les prédictions des structures sur le protéome complet de l'espèce *Methanocaldococcus jannaschii* prennent 242 heures au maximum contre les 4200 heures de calcules estimées avec Alphafold2 [40].

5.4.3 . Prédiction des PPI en utilisant Colabfold-Multimer

Colabfold permet d'atteindre des performances proches de celle d'Alphafold2 et Alphafold-Multimer pour la prédiction de structure, mais en ayant une vitesse d'exécution plus rapide. Cependant qu'en est-il du problème de prédiction de PPI ? L'utilisation de l'implémentation d'Alphafold-Multimer par Colabfold (Colabfold-Multimer) pourrait peut être permettre d'avoir une méthode de prédiction de PPI encore plus performante que celle de l'article de Humphrey et al [20], et la vitesse de calcul accélérée par colabfold pourrait peut être permettre de se passer de Rosettafold, certes très rapide mais qui pourrait produire de nombreux faux-négatifs.

Pour répondre à cette question, j'ai donc rassemblé un set de PPI positif et négatif sur lequel j'ai lancé Colabfold-Multimer.

5.4.3.1 . Le set de PPI Gold Standard (GS) de *S. cerevisiae*

Pour évaluer les performances de Colabfold-Multimer sur la prédiction de PPI, j'ai récupéré le jeu de données de PPI Gold Standard utilisé dans le papier [20]. Cela représente un total de 548 PPI chez *S. cerevisiae* impliquant 589 protéines. Les résultats de prédiction par Rosettafold+Alphafold et par ColabFold peuvent ainsi être comparés pour ces 548 PPI.

5.4.3.2 . Le set de PPI négatif de *S. cerevisiae*

Pour construire un set de PPI négatif, j'ai choisi de reprendre la méthodologie la plus couramment utilisée dans les articles de prédiction de PPI et qui consiste à sélectionner aléatoirement des paires de protéines.

Tout d'abord, j'ai récupéré l'ensemble des PPI de *S. cerevisiae* présentes

dans les bases de données Biogrid [48], Intact [47] et Irefindex [55] que j'ai fusionnées en utilisant APPINetwork. Ce fichier me permet de vérifier par la suite que les paires de protéines choisies aléatoirement ne sont pas des PPI.

J'ai ainsi généré 110 paires aléatoires impliquant des protéines du jeu de données Gold Standard, auxquelles j'ai ajouté 77 paires aléatoires à partir des 467 protéines de *S. cerevisiae* pour lesquelles les cartes IPOPS ont été générées. Enfin j'y ai rajouté 410 paires aléatoires choisies dans le protéome de *S. cerevisiae*. Cela fait un total de 597 paires de protéines pour le set de négatif.

Ce jeu de données ne reflète pas la réalité puisqu'il n'est pas aussi déséquilibré que celui d'Hymphrey et al. Mes résultats ne seront donc pas comparables à ceux de leur papier. Par contre, je pourrai comparer sur ce jeu de données les résultats avec l'approche "RosettaFold+AlphaFold2" à l'approche "ColabFold-multimer"

5.4.3.3 . Le set de PPI Gold Standart et le set négatifs d' *A. thaliana*

J'ai récupéré un set de 119 PPI de confiance, déjà utilisées dans un autre papier [8]. Ce set a été créé en cherchant dans la littérature les PPI de confiance décrites par au moins 2 papiers différents et 2 méthodes différentes.

J'ai créé un set de 300 PPI aléatoires pour le set de négatif en créant des paires aléatoires parmi l'ensemble des protéines d'*A. thaliana* et en vérifiant qu'elles étaient bien absentes des bases de données Intact, Biogrid, Irefindex et TAIR (une base de données spécialisée pour l'organisme *A. thaliana*) [52].

5.4.3.4 . Les paramètres de Colabfold-Multimer pour la prédiction des complexes

Les options suivantes ont été utilisées pour prédire la structure des complexes de l'ensemble des paires de protéines du jeu de gold standard et du jeu de négatifs avec Colabfold-Multimer :

- les pMSA ont été générées sur le serveur public MMSeqs2 mis en place par Colabfold.
- –save-all : sauvegarde toutes les sorties intermédiaires générées par Colabfold-Multimer. Cette option est nécessaire afin de récupérer les matrices de distances nécessaires à la prédiction de PPI.
- –num-recycle : le nombre de recyclages maximum autorisés pour la prédiction de la structure 3D du complexe est 20 et le nombre par défaut est 3. Nous avons donc testé ces deux valeurs.
- –num-models : 5 modèles sont générés par Colabfold-Multimer pour une paire de protéines. Réduire le nombre de modèles per-

mettrait de réduire le temps de calcul. Cependant j'ai fait le choix de laisser ce paramètre sur 5 afin de déterminer le nombre de modèles permettant de maximiser les performances de la prédiction de PPI.
- –rank : permet de choisir la métrique qui sera utilisée pour classer les 5 modèles. J'ai utilisé l'ipTM car c'est le paramètre utilisé par défaut dans Colabfold-Multimer.

5.4.3.5 . Prédiction des PPI à partir des sorties de Colabfold-Multimer

Conformément au protocole décrit dans [20], pour prédire des PPI à partir des prédiction de Colabfold multimer, j'ai extrait des sorties brutes, les matrices de distances décrites dans 5.3.2. Ces matrices ne donnent pas directement les probabilités pour les paires d'acides aminés de se trouver dans les différents bins de distances, mais plutôt des logits. Le logit est une valeur numérique qui représente la confiance qu'a le modèle qu'une paire d'acides aminés appartienne à un bin de distances. J'ai donc utilisé la fonction softmax pour transformer les logits en probabilités, en s'assurant que la somme de toutes les probabilités pour l'ensemble des bins de distance est égale à 1

$$\text{softmax}(z)_i = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}$$

où z est le vecteur de logits, z_i est le logit pour le bin de distance i, et K est le nombre total de bins. Pour l'ensemble des paires d'acides aminés, je calcule la probabilité qu'une paire se trouve à moins de 12 Åen sommant les probabilités des bins de distances jusqu'à 12 Å. Enfin la probabilité de contact que j'utilise pour la prédiction de PPI est la probabilité de contact maximum sur l'ensemble des paires d'acides aminés n'appartenant pas à la même chaîne protéique.

Dans ce protocole décrit par [20], une PPI est prédite à partir d' une seule probabilité de contact. Il pourrait être intéressant d'utiliser plusieurs probabilités de contact pour prédire l'interaction mais l'hypothèse dans notre protocole est que la probabilité de contact maximum dépend déjà de l'environnement de la paire d'acides aminés de probabilité maximale et donc prend déjà en compte plusieurs probabilités de contact.

6 - Résultats et discussion

6.1 . Prédiction chez *S. cerevisiae*

6.1.1 . Résultats de la prédiction utilisant Colabfold-Multimer et ses paramères par défaut

6.1.1.1 . Performances de la prédiction

J'ai évalué les performances de la méthodes de prédiction de PPI sur les sets négatif et positif de *S. cerevisiae* (voir section 5.4.3) en construisant des courbes représentant la précision de la méthode en fonction du recall lorsque le seuil de décision sur la probabilité de contact diminue. Cette courbe permet d'évaluer la capacité du modèle à prédire les PPI de manière fiable et à ne pas manquer de PPI pour chaque valeur du seuil de décision. Cela permet notamment de choisir un seuil de décision pour la classification. Il y a en effet toujours un compromis à trouver entre précision et recall, certaines application demande de prédire les exemples positifs avec une grande confiance, d'autres nécessitent de ne louper aucune prédiction positive au détriment de la précision.

Lorsqu'on s'intéresse à un interactome complet, Le nombre de paires de protéines à prédire est très élevé. Cependant, les instances positives sont des évènements rares et la chance de tomber aléatoirement sur une paire de protéines capable d'interagir de manière fonctionnelle est donc relativement faible. Le compromis à réaliser est donc généralement en faveur de la précision. Dans le papier [20], le seuil sur la probabilité de contact a été choisi de façon à garantir une précision de 0.95 pour un recall de 0.29.

Sur la figure 6.1, pour la classification en PPI et non PPI, j'ai utilisé la probabilité de contact maximum du modèle avec l'ipTM score le plus élevé. Pour garantir une précision de 0.95, le seuil sur la probabilité de contact doit être fixé 0.964. Pour cette valeur de seuil, le recall est de 0.55 (voir figure 6.1).

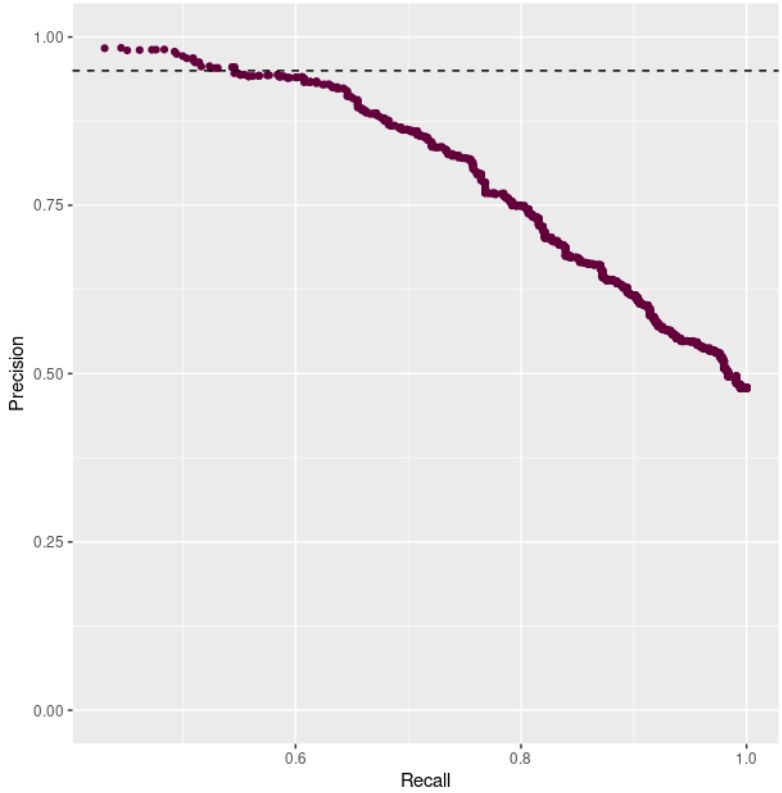


Figure 6.1 – **Performance de Colabfold-Multimer pour la prédiction de PPI en prenant seulement en compte la structure avec le meilleur ipTM.**

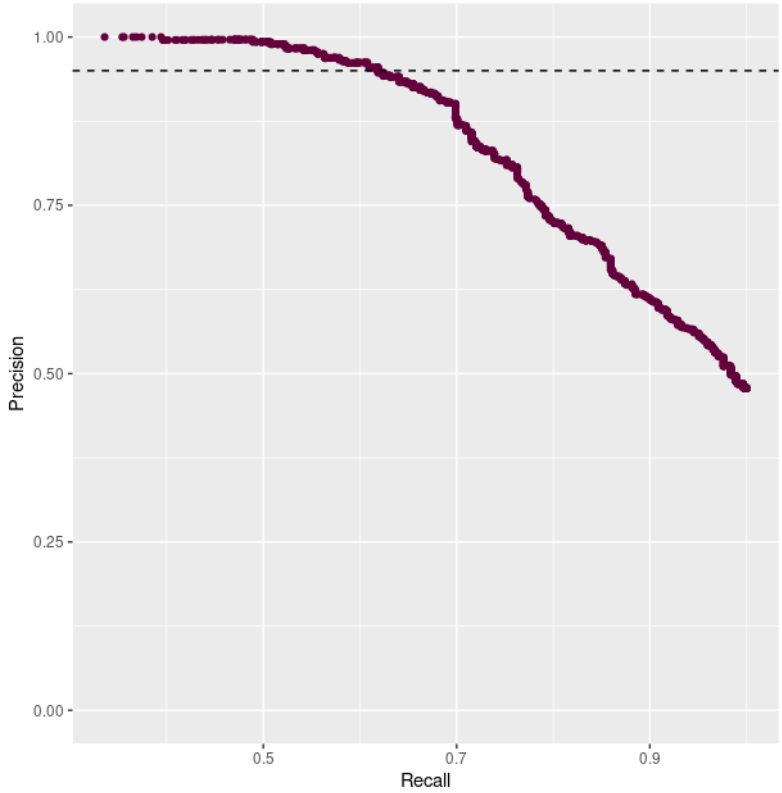


Figure 6.2 – **Performance de Colabfold-Multimer pour la prédiction de PPI en prenant seulement en compte la moyenne de la probabilité de contact maximum sur les 5 modèles.**

Le seuil fixé sur la probabilité de contact semble particulièrement élevé. Sur les paires de protéines sélectionnées aléatoirement, Alphafold-Multimer trouve donc des structures de complexes avec des probabilités de contact élevées. Cela pourrait donc faire écho à des papiers publiés précédemment qui montrent que pour une proportion non négligeable de paires de protéines sélectionnées aléatoirement, il n'existe pas d'incompatibilité notable au niveau de la structure 3d des protéines rendant impossible l'interaction [30]. Pour mieux différencier les paires formant des interactions fonctionnelles des paires formant des interactions non-fonctionnelles, il faudrait éventuellement se tourner vers d'autres métriques. Peut-être qu'une interface a été trouvée par Alphafold-Multimer mais que la qualité estimée localement le long de chacune des chaînes (par le pLDDT) n'est pas bonne ? Peut-être que l'organisation des domaines les un par rapport aux autres (estimée à partir de la matrice de PAE) n'est pas correcte ? Je n'ai pas eu le temps durant ma thèse d'explorer ces possibilités sur mon jeu de données prédits mais toutes ses informations sont disponibles et pourront faire l'objet d'une analyse intéressante. Les structures de l'ensemble des complexes prédits sont aussi disponibles. Ainsi il sera intéressant d'étudier les caractéristiques des interfaces trouvées par Alphafold-Multimer, notamment leurs différentes caractéristiques physico-chimiques ainsi que leur propension à l'interaction calculée grâce au pipeline IPOPS.

6.1.1.2 . Exemples positifs dans le set de négatifs

La probabilité de tomber sur une PPI lorsque des paires de protéines sont sélectionnées de manière aléatoire est normalement très faible. Cependant parmi les paires aléatoires de mon jeu de données dont la probabilité de contact était très élevée (supérieure au seuil de décision) et donc considérées comme faux positifs, je suis malgré tout tombé sur des exemples intriguant.

- La paire de protéines du set aléatoire (Q08749 ; P21801) est prédite comme une PPI par la méthode avec un score de 0.91 (moyenne des 5 modèles). Cette paire n'est pas renseignée dans l'ensemble des bases Irefindex, Biogrid et Intact. Or, lorsque l'on regarde les informations disponibles sur Uniprot, on remarque que la fonction de Q08749 est liée à l'insertion de protéines transmembranaires dans la membrane interne de la mitochondrie et que P21801 est une protéine transmembranaire localisée dans la membrane interne de la mitochondrie.

- La paire de protéines du set aléatoire (P00044 ; P00410) est prédite comme une PPI avec un score de 0.98. P00044 est une isoforme

du cytochrome C qui joue un rôle dans la respiration cellulaire, et notamment dans le transport des électrons vers la sous unité COX2 de la cytochrome oxydase et justement P00410 est COX2. Cependant l'interaction n'est pas présente dans les bases de données. L'interaction entre le cytochrome c et COX2 a cependant été décrite, notamment chez les mammifères. [68].

- La paire de protéine du set aléatoire (P02992; P0CX31) est prédite comme une PPI avec un score de 0.88. P02992 est une protéine chargée de transporter les aminoacyls ARNt, nécessaires pour la synthèse d'une chaîne protéique par le ribosome. P0CX31 est la petite sous unité du ribosome.

Ces trois paires de protéines sont tirées du set négatif. Bien que cela illustre la capacité de la méthode à identifier des PPI, ce résultat est surprenant. Cela pourrait indiquer qu'il existe un biais fonctionnel ou un biais sur la localisation cellulaire dans les 110 paires choisies aléatoirement à partir des protéines gold standart, et plus largement dans le set Gold Standard ou dans les 77 paires choisies aléatoirement à partir des 467 protéines de *S. cerevisiae* pour lesquelles les cartes IPOPS ont été générées.

6.1.1.3 . Cartes IPOPS et interfaces prédites par ColabFold

J'ai ensuite voulu regarder avec un exemple s'il était possible d'éxécuter le pipeline IPOPS sur un modèle généré par Alphafold2 et si les îlots trouvés avec le pipeline se retrouvaient au niveau des interfaces trouvées par Colabfold-Multimer. Pour cela j'ai sélectionné un complexe du set de gold standard avec une probabilité de contacts élevé, (P06242; P37366). J'ai récupéré le modèle prédit par Alphafold2 d'une des deux protéines du complexe, P06242. J'ai docké le modèle Alphafold2 avec 100 protéines d'un set que l'on appelle le set de Background. Enfin, j'ai lancé le pipeline IPOPS à partir de ces dockings, j'ai extrait les acides aminés appartenant aux îlots rouges et roses (pour rappel 4.2.5) pour les afficher sur la structure du complexe prédite par Colabfold-Multimer. Sur cet exemple, les acides aminés appartenant aux îlots rouges et roses semblent bien se retrouver au niveau de l'interface protéique. Il sera intéressant dans la suite du projet de regarder sur plus d'exemples les propriétés des cartes IPOPS aux interfaces prédites.

Sur les 548 PPI du set de Gold Standard, seules 29 PPI ont des cartes IPOPS déjà générées pour les deux partenaires de la paire (voir 4.3.1). Pour compléter ce set, j'ai donc généré des dockings avec ATTRACT avec 17 PPI supplémentaires, en partant des modèles des monomères extraits d'AlphafoldDB.

Dans le set négatif, il y a 77 paires pour lesquelles j'ai déjà généré les cartes IPOPS.

Les cartes des propriétés physico-chimiques ont été calculées pour l'ensemble des protéines du set Gold standard et du set négatif.

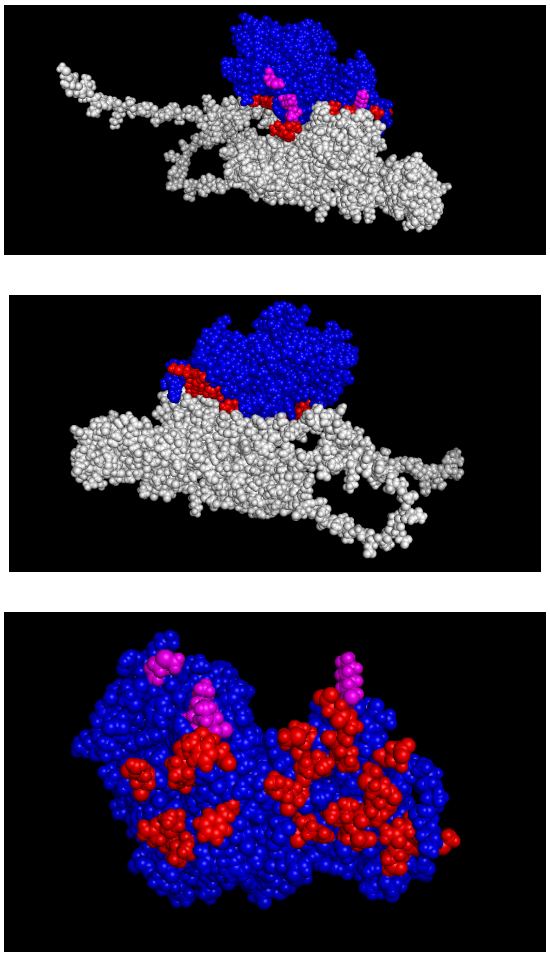


Figure 6.3 – **Présence des îlots IPOPS à l'interface du complexe (P06242 ; P37366) prédits par Colabfold-Multimer.** La protéine en bleu est P06242. Les acides aminés sont en rouges sont ceux qui, sur les cartes IPOPS, appartiennent aux îlots rouges (dont l'interaction est favorable à un grand nombre de ligands) ou roses (dont l'interaction est favorable à un plus faible nombre de ligands).

6.1.2 . Réduction du temps de calcul

6.1.2.1 . Les différentes options permettant de réduire le

temps de calcul

La durée moyenne d'une prédiction est beaucoup trop élevée pour envisager de prédire un très grand nombre de PPI. J'ai donc exploré plusieurs approches permettant de réduire le temps de calcul de la prédiction de PPI sans trop sacrifier les performances. Pour cela, plusieurs options s'offraient à moi :

- la première option consistait à établir un seuil sur la métrique de qualité (ici l'ipTM) utilisée pour classer les modèles. Lors de la prédiction d'un modèle, si à la suite de plusieurs étapes de recyclages ce seuil de qualité est atteint, aucune autre étape de recyclage supplémentaire n'est réalisée.

- La seconde option consistait à choisir un nombre de recyclages maximum. Par défaut, le nombre de recyclages maximum autorisés lors de la prédiction de complexe avec Alphafold-multimer est de 20.

- La troisième option était de diminuer le nombre de modèles produits pour une paire de protéines.

- La quatrième option consistait à utiliser une méthode dont le temps de calcul est plus faible pour présélectionner les PPI les plus probables et ainsi fournir un nombre réduit de PPI à Colabfold-Multimer, comme cela a été fait avec Rosetta2track et Alphafold2 dans [20].

Mon temps étant limité pour produire des résultats, nous n'avons pas exploré la première option. Par contre, nous avons exploré les trois autres.

6.1.2.2 . Prédiction sur un jeu de données plus restreint

Pour étudier l'impact de ces différentes options sur le temps de calcul et sur les performances, j'ai utilisé un même jeu de données qui est un peu différent de celui décrit dans la section 5.4.3. En effet, pour tester les performances de Rosetta2track, je l'ai lancé sur :

- l'ensemble des 548 paires du set positif. Mais les calculs n'ont abouti que pour 425 d'entre elles.

- l'ensemble des 589 paires du set négatif. Mais les calculs n'ont abouti que pour 407 paires.

Si les calculs n'ont pas abouti pour certaines paires de protéines c'est que Rosetta2track ne permet pas de prédire des complexes aussi grand que Colabfold-Multimer. La taille maximale autorisée dépend de la mémoire graphique dédiée du GPU utilisé. Pour la RTX A5OOO, cette limite se situe aux alentour de 1300 aa.

Afin d'avoir un jeu de données équilibré, j'ai sous échantillonné le set positif de façon à avoir 407 paires dans chacun des deux sets (positif et négatif). Dans les paragraphes qui suivent, l'ensemble des performances seront évaluées sur ce set équilibré.

6.1.2.3 . Impact du nombre de recyclages sur le temps de calcul et sur les performances de la prédiction de PPI

Dans la littérature, on trouve une étude sur l'impact du nombre de recyclages sur les prédictions de structures. Ainsi, dans [5], le fait d'utiliser 10 recyclages ou 3 recyclages ne semble pas avoir d'impact majeur sur le nombre de modèles considérés comme acceptables. Mais on peut se demander si c'est aussi le cas pour la probabilité de contact maximum extraite des distogrammes produits par Alphafold-multimer.

Afin d'étudier l'impact du nombre de recyclages sur les performances de la prédiction de PPI, j'ai lancé les prédictions sur l'ensemble des PPI du set décrit pécédemment (en section 6.1.2.1) avec 3 et 20 recyclages.

Concernant le gain en terme de temps de calculs pour la prédiction d'un complexe de mille acides aminés, le temps de calcul pour 5 modèles est de 1h18 en moyenne pour 20 recyclages contre 14 minutes en moyenne pour 3 recyclages sur une NVDIA RTX A5000 disposant de 24 Go GDDR6. La réduction du nombre de recyclage est donc très efficace pour réduire le temps de calcul. Mais impacte-t-elle les performances de prédiction des PPI ?

La réponse vient des résultats fournis dans le tableau 6.1. On remarque en effet qu'avec 20 recyclages, le recall obtenu est plus faible qu'avec 3 recyclages lorsqu'on utilise la probabilité de contact maximale du meilleur modèle ou la moyenne des probabilités de contact de 2 ou de 3 modèles. On comprend pourquoi en regardant les distributions des probabilité de contact des figures 6.5 et 6.6. En effet, on s'aperçoit que la distribution des probabilités des paires aléatoires de protéines (courbe rouge) est plus décalée vers la gauche lorsqu'on utilise 3 recyclages que 20 recyclages. Comme la distribution des probabilités des PPIs (courbe bleue) change peu entre 3 et 20 recyclages (sauf quand on utilise un seul modèle), les distributions du set positif et du set négatif sont mieux sépa-

rées et le résultat de la prédiction n'en n'est que meilleur. On voit d'ailleurs dans le tableau 6.1 que le seuil sur la probabilité de contact utilisé pour la prédiction est plus faible avec 3 recyclages. En effet, les 20 recyclages affinent la prédiction de la structure des modèles et augmentent leur probabilité de contact même pour des paires de protéines qui n'interagissent pas.

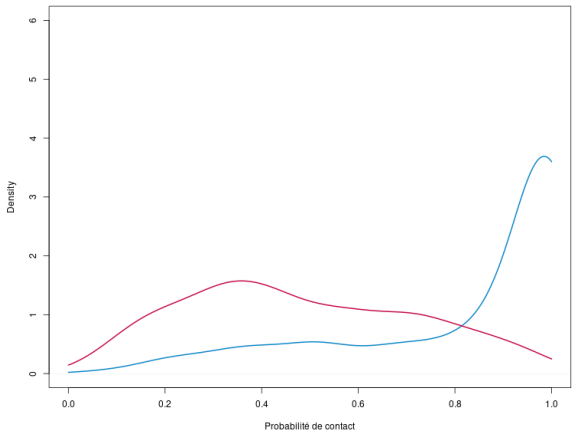
Par ailleurs, sur la figure 6.6 les courbes représentant la précision en fonction recall pour les 4 meilleurs modèles et pour 3 (figure *a*) et 20 (figure *b*) recyclages, montrent que la la courbe de la figure *a* présente un plateau initial plus grand que la courbe de la figure *b* où la précision diminue plus progressivement. Cela signifie que la distribution des probabilités du set négatif chevauche moins celle du set positif, en tous cas au niveau des probabilités élevées. Si l'on était plus strict et que l'on cherchait à avoir une précision encore plus élevé, la prédiction avec l'option 3 recyclages serait donc plus intéressante.

Ces résultats sont intéressants car ils montrent ce que nous voulions voir, c'est à dire qu'on peut réduire le nombre de cycles sans diminuer les performances de la prédiction, bien au contraire. On peut même se demander si le recyclage est nécessaire.

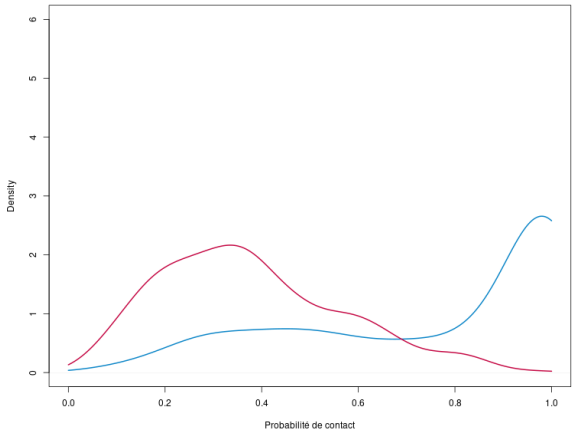
A	3 recyclages				
	1 modèle	2 modèles	3 modèles	4 modèles	5 modèles
Seuil	0.928	0.885	0.876	0.826	0.807
Précision	0.95	0.95	0.95	0.95	0.95
Recall	0.57	0.58	0.57	0.59	0.58

B	20 recyclages				
	1 modèle	2 modèles	3 modèles	4 modèles	5 modèles
Seuil	0.975	0.929	0.882	0.838	0.828
Précision	0.95	0.95	0.95	0.95	0.95
Recall	0.50	0.56	0.60	0.62	0.59

Table 6.1 – **Tableaux présentant les performances de la prédiction de PPI avec 3 (A) et 20 (B) recyclages maximum et avec un ou plusieurs modèles**. Le seuil sur la probabilité de contact utilisé pour obtenir une précision qui soit égale à 0.95 est donné sur la première ligne et le Recall correspondant est fourni à la dernière ligne du tableau. Le nombre de modèles utilisés pour le calcul de la probabilité de contact varie au niveau des colonnes.

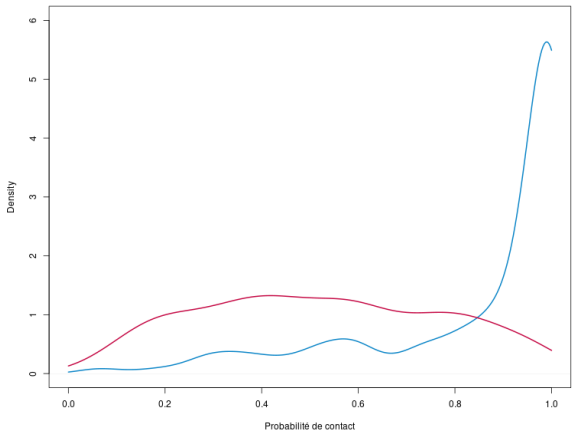


(a)

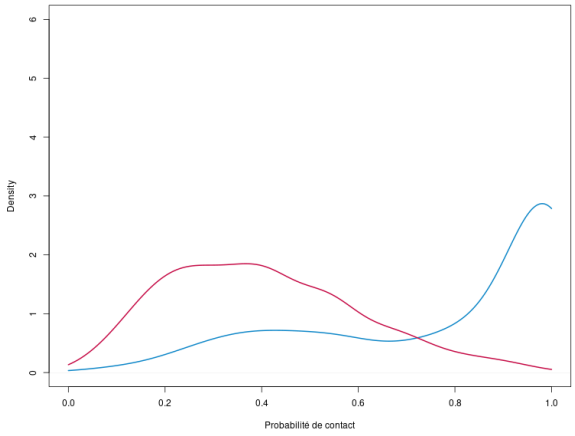


(b)

Figure 6.4 – **Distribution des probabilités de contact pour 1 modèle (a) et 5 modèles (b) pour les négatifs (en rouge) et les positifs (en bleu) lorsque les modèles sont générés avec 3 recyclages maximums**

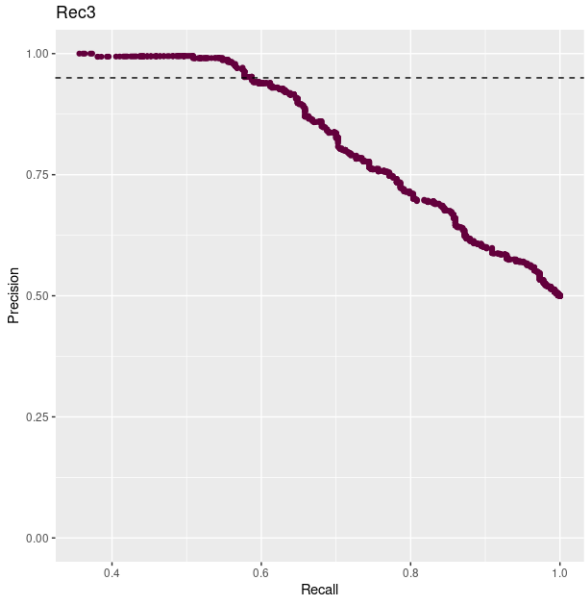


(a)

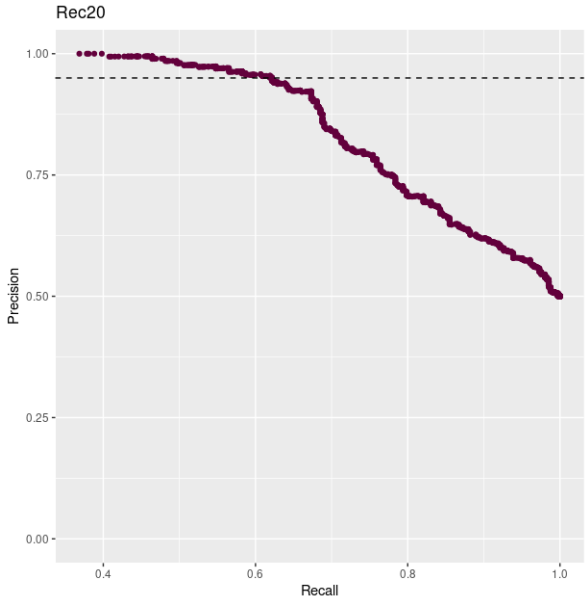


(b)

Figure 6.5 – **Distribution des probabilités de contact pour 1 modèle (a) et 5 modèles (b) pour les négatifs (en rouge) et les positifs (en bleu) lorsque les modèles sont générés avec 20 recyclages maximum**



(a)



(b)

Figure 6.6 – **Performances de Colabfold-Multimer lorsque la moyenne des probabilités de contact des 4 meilleurs modèles est utilisée en autorisant 3 recyclages (a) et 20 recyclages (b) au maximum.**

6.1.2.4 . Impact du nombre de modèles sur les performances de la prédiction de PPI

La réduction du nombre de modèles impacte-t-elle les performances de prédiction des PPI ? Pour répondre à cette question, regardons dans le tableau 6.1 les résultats des performances en fonction du nombre de modèles utilisés pour calculer les probabilités de contact. On voit que quelque soit le nombre de recyclages, les meilleures performances sont atteintes lorsque l'ont fait la moyenne des probabilités de contact des 4 meilleurs modèles. Ce résultat est cependant à relativiser, pour 3 recyclages car le recall ne change quasiment pas entre les prédiction utilisant 2, 3, 4 ou 5 modèles pour calculer les probabilité de contact. Pour 20 recyclages, le gain sur le recall est beaucoup plus important quand on passe de 1 modèle à 4 modèles (0.12). L'écart de performance semble donc négligeable par rapport au gain de temps important obtenu par la réduction du nombre de recyclages.

On remarque par ailleurs, que lorsque le nombre de modèles utilisés augmente, le seuil de probabilité correspondant à 0.95 de précision diminue. En effet, lorsqu'on utilise un plus grand nombre de modèles, qui sont moins bons que le premier selon l'ipTM, la probabilité de contact diminue. C'est ce qu'on peut voir sur les figures 6.4 et 6.5 où les distributions des probabilités de contact de la figure du bas (5 modèles) sont plus décalées vers la gauche (vers les probabilités plus faibles) que la figure du haut (1 modèle). Une autre observation intéressante sur ces courbes est que le nombre de modèles affecte peu l'allure de la distribution des probabilités de contact du set positif (sauf pour 1 modèle) alors qu'il impacte fortement celle du set négatif qui est se décale vers la gauche. L'augmentation du nombre de modèles utilisés permet ainsi une meilleure séparation des distributions des probabilités de contact du set négatif et du set positif et donc une meilleure prédiction.

Ces résultats vont donc dans le sens du maintient du calcul avec 5 modèles, d'autant plus si on utilise 3 recyclages.

6.1.2.5 . Utilisation de Rosettafold en tant que filtre pour Colabfold-multimer

N'ayant pas de set aussi déséquilibré que le set de l'article de Humphrey et al [20], la comparaison directe de mes performances avec les leurs n'est pas possible. Cependant, il est possible d'évaluer la contribution de rosettafold au niveau des performances de prédiction pour savoir si son utilisation est intéressante en tant que filtre pour Colabfold-Multimer.

Pour cela, j'ai lancé Rosetta2track [20] sur le set décrit dans la section 5.4.3. Comme je l'ai écrit précédemment en section 6.1.2.2, pour cause de

limitation de taille de séquence, seuls 425 calculs ont abouti sur le set positif et 407 sur le set négatif, .

Rosetta2track fournit directement la probabilité de contact maximum. De la même manière que pour Colabfold, cette probabilité de contact est utilisée pour prédire des PPI et les performances sont évaluées sur le set équilibré décrit en section 6.1.2.2.

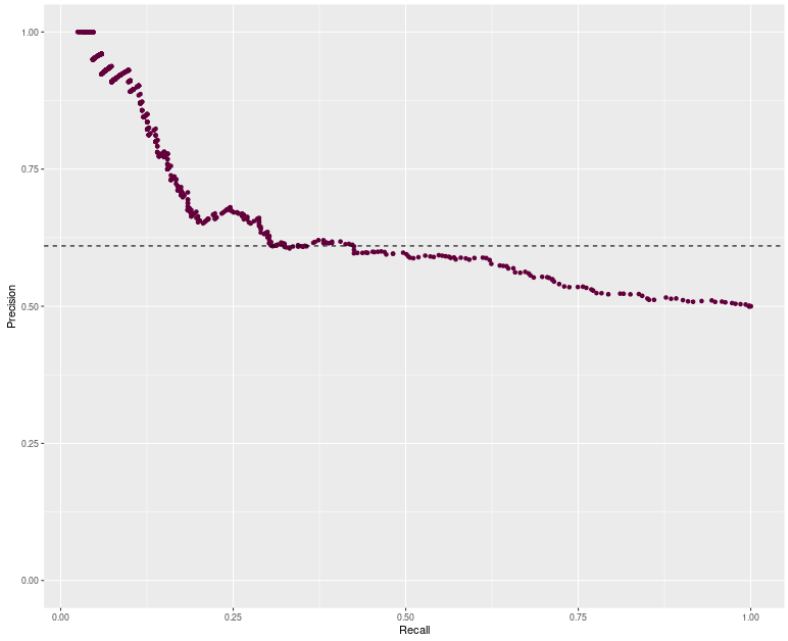


Figure 6.7 – **Performances de Rosetta2track pour la prédiction de PPI**

Choisir un seuil sur Rosetta2track permet de filtrer plus où moins de paires de protéines. Un seuil plus élevé permet de filtrer plus de paires de protéines, mais au prix d'une diminution du recall, ce qui se ressentira sur les résultats combinés de Rosetta+Colabfold. J'ai choisi de placer le seuil à une probabilité de contact de 0.095 ce qui correspond à une précision de 0.61 et un recall de 0.43 (voir figure 6.7. Cela permet de filtrer 65% des paires de protéines du set équilibré. La prédiction de PPI est ensuite réalisée sur les 35% de paires restantes avec Colabfold-Multimer 3 recyclages en prenant en compte les 4 meilleurs modèles.

Sur la figure 6.8, la courbe représentant la précision en fonction du recall, obtenue en combinant RosettaFold et ColabFold, présente un plateau initial, puis les performances diminuent progressivement. Au seuil de précision de 0.95, le recall est de 0.30. C'est donc une perte de 0.29 de recall par rapport aux performances de Colabfold seules.

Concernant le temps d'exécution, Rosetta2track réalise ses prédictions en 44 secondes en moyenne pour une paire de protéines. Cela est beaucoup plus court que le temps de calcul de Colabfold-Multimer qui est de

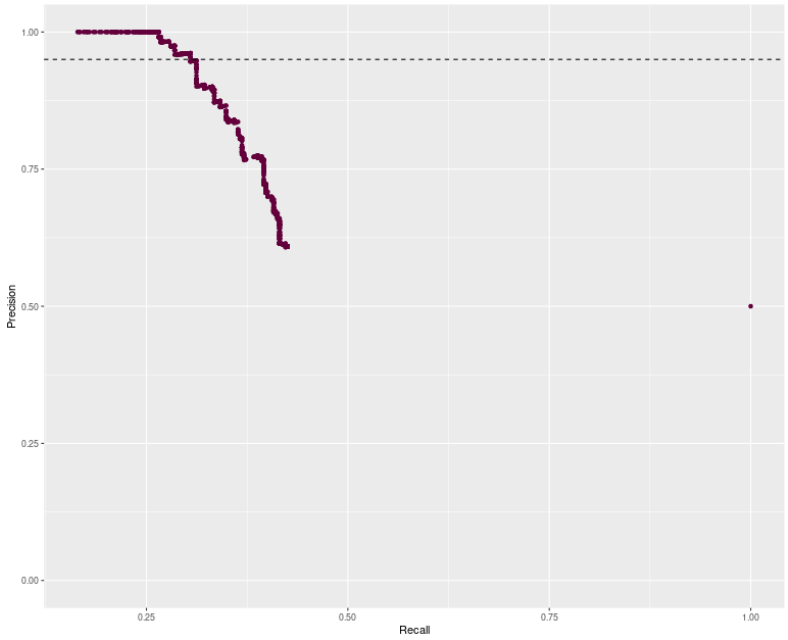


Figure 6.8 – **Performances de Rosetta+Colabfold-Multimer pour la prédiction de PPI.** La probabilité de contact des PPI filtrées a été arbitrairement mise à 0 pour les prédictions avec Colabfold-Multimer.

14 minutes en moyenne. Cependant, il faut prendre en compte le fait que, pour chaque paire de protéines à prédire, il faut réaliser la construction de la MSA au préalable. Cette étape prend en moyenne 4 minutes 30 sur le serveur publique MMSEQS2. Le seuil de Rosetta2track a été placé de façon à filtrer 65% du jeu de données. Dans 65% des cas, la durée du pipeline est de 5 minutes 14 car la prédiction par ColabFold n'est pas réalisée. Dans 35% des cas, la durée du pipeline est de 5 minutes 14 + 14 minutes pour la prédiction avec Colabfold (l'alignement n'est réalisé qu'une seul fois). En moyenne, une prédiction avec Rosetta+Colabfold prend donc : $5'14 \times 0.65 + 0.35 \times (5'14 + 14') = 608$ secondes. Concernant Colabfold seul le temps de calculs est de 18 minutes 30 en moyenne pour la prédiction et l'alignement soit 1110 secondes. Le gain de temps apporté par Rosetta2track est donc modéré par rapport à la baisse de performances observée.

6.1.2.6 . Performances sur un set déséquilibré

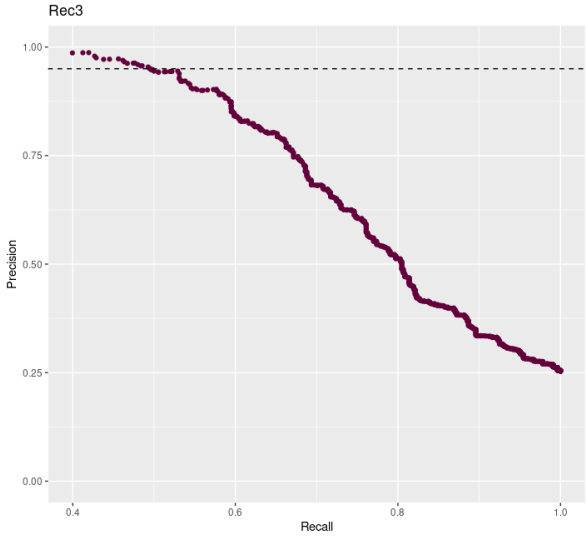
Durant les derniers mois de ma thèse, j'ai demandé un accès au supercalculateur Jean Zay. J'ai lancé des alignements de séquences pour un peu moins de 2000 paires aléatoires. Je n'avais pas encore les résultats complets de Rosetta2track, mais nous avions déjà remarqué que les prédictions pour des paires d'environ 1300 aa n'aboutissaient pas. Les

alignements générés ne concernent que des paires dont la taille combinée est inférieure à 1300 aa, au cas où nous adopterions la stratégie Rosetta2track+ColabFold-Multimer. Finalement, au vu des résultats présentés dans les paragraphes précédents, la stratégie adoptée a été d'utiliser ColabFold-Multimer 3 recyclages 5 modèles pour prédire plus de paires aléatoires et obtenir des sets plus déséquilibrés. Le set final chez *S. cerevisiae* comprend 548 paires gold standards et 1612 paires aléatoires. Les performances sur ce set déséquilibré confirment les résultats obtenus précédemment, le fait d'utiliser plusieurs modèles permet de mieux distinguer les paires positives des paires aléatoires. À une précision de 0.95, les meilleures performances sont obtenues lorsque les 5 modèles sont pris en compte, avec un recall de 0.581 au seuil de 0.835. Concernant les courbes de précision/recall, on observe un plateau initial d'autant plus important que le nombre de modèles pris en compte est élevé (voir Figure 6.9).

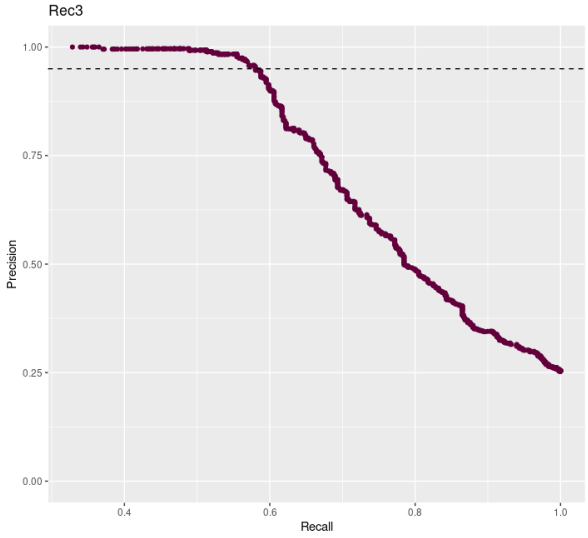
Pour comparer les performances de la méthode que j'ai mise en place à celle de Humphreys et al. [20], pour chaque valeur de seuil sur la probabilité de contact, j'ai calculé le FPR à partir des 548 PPI gold standards et des 1612 paires aléatoires, puis je l'ai multiplié par 548 000 afin d'obtenir un nombre de faux positifs estimé dans le cas d'un déséquilibre similaire à celui de [20]. Ce nombre de FP est utilisé pour recalculer la précision. J'ai également réalisé un intervalle de confiance à 95% de façon à pouvoir visualiser la plage de valeurs que pourrait prendre la précision ainsi estimée. Lorsque l'on regarde la figure 6.10, la plage de valeurs que peut prendre la précision, est très large, car l'ensemble à partir duquel nous avons réalisé notre estimation est très loin du déséquilibre de 1 positif pour 1000 négatifs. Cependant, en s'intéressant au pire des cas possibles représenté par la courbe bleue, le recall est de 0.36 pour 1 de précision avant que celle-ci ne diminue. La méthode mise en place par Humphreys et al dans [20] a un recall de 0.29 pour 0.95 de précision. Donc dans le pire cas possible, la méthode que j'ai mise en place présente de meilleures performances pour une augmentation modérée du temps de calcul comme montré dans le paragraphe 6.1.2.5.

A	3 recyclages				
	1 modèle	2 modèles	3 modèles	4 modèles	5 modèles
Seuil	0.972	0.933	0.9	0.865	0.835
Précision	0.95	0.95	0.95	0.95	0.95
Recall	0.495	0.553	0.568	0.577	0.581

Table 6.2 – **Tableaux présentant les performances de la prédiction de PPI de Colabfold-Multimer 3 recyclages et avec un ou plusieurs modèles**. Le seuil sur la probabilité de contact utilisé pour obtenir une précision qui soit égale à 0.95 est donné sur la première ligne et le Recall correspondant est fourni à la dernière ligne du tableau. Le nombre de modèles utilisés pour le calcul de la probabilité de contact varie au niveau des colonnes.



(a)



(b)

Figure 6.9 – **Performances de Colabfold-Multimer 3 recyclage sur un set déséquilibré lorsque le meilleur modèle est pris en compte (a) ou les 5 meilleurs modèles sont pris en compte (b)**

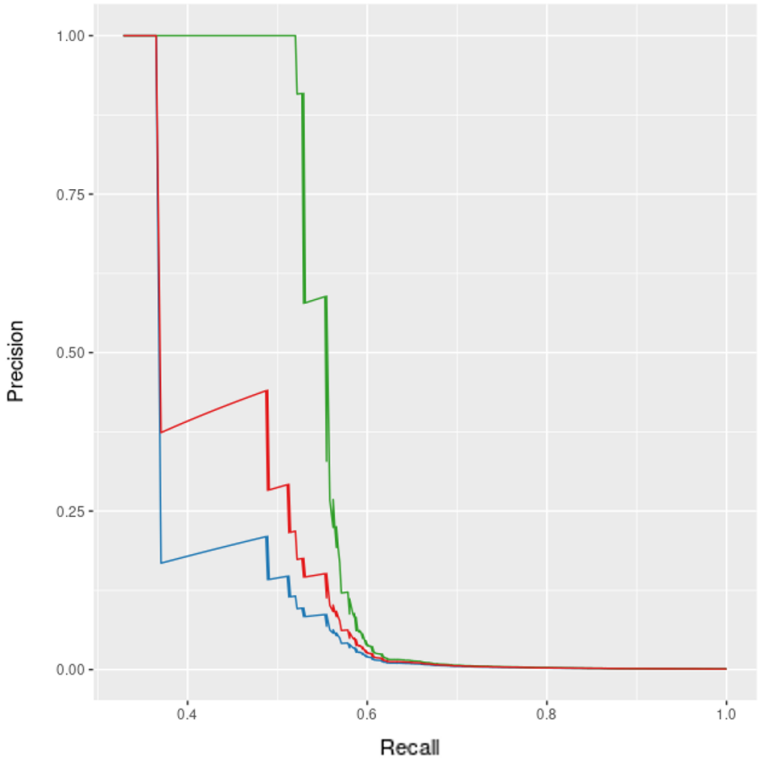


Figure 6.10 – **Performances de Colabfold-Multimer 3 recyclage 5 modèles estimées lorsque la taille de l'ensemble négatif est de 548 000 paires aléatoires (courbe rouge). Les courbes bleues et vertes représentent les bornes supérieures et inférieures de l'intervalle de confiance à 95% calculées pour le False Positive Rate**

6.1.3 . Utilisation de cette approche chez *A. thaliana*

Enfin, j'ai cherché à voir si Colabfold-Multimer permettait de prédire des PPI avec des performances acceptables chez *A. thaliana*. En effet, l'objectif initial de ma thèse était d'explorer l'interactome de cet organisme modèle, et chez les plantes, il y a de nombreux évènements de duplication de gènes ce qui peut rendre l'identification d'orthologues plus difficile. Cela aurait pour conséquence l'obtention de MSA plus bruitées et donc une diminution des performances.

Suites aux résultats obtenus chez *S. cerevisiae* avec les différents paramètres si dessus, j'ai lancé des prédictions en utilisant Colabfold-Multimer avec comme option 3 recyclages et 5 modèles sur le jeu de données d'*A. thaliana* décrit section 5.4.3.3.

A 0.95 de précision, les meilleures performances sont observées lorsque 4 modèles sont pris en compte. Le seuil au niveau de la probabilité de contact est de 0.898 et le recall est de 0.56. Les performances ne semble donc pas bien différentes de celles observées chez *S. cerevisiae*, d'autant que le nombre paires du set aléatoire est le double de celui du set gold standard.

Cependant le seuil de probabilité de contact est différent du celui chez *S. cerevisiae*. Il sera donc nécessaire d'étoffer le set positif et négatif chez *A. thaliana* afin de pouvoir fixer un seuil de probabilité plus représentatif, permettant de prédire des paires chez *A. thaliana*.

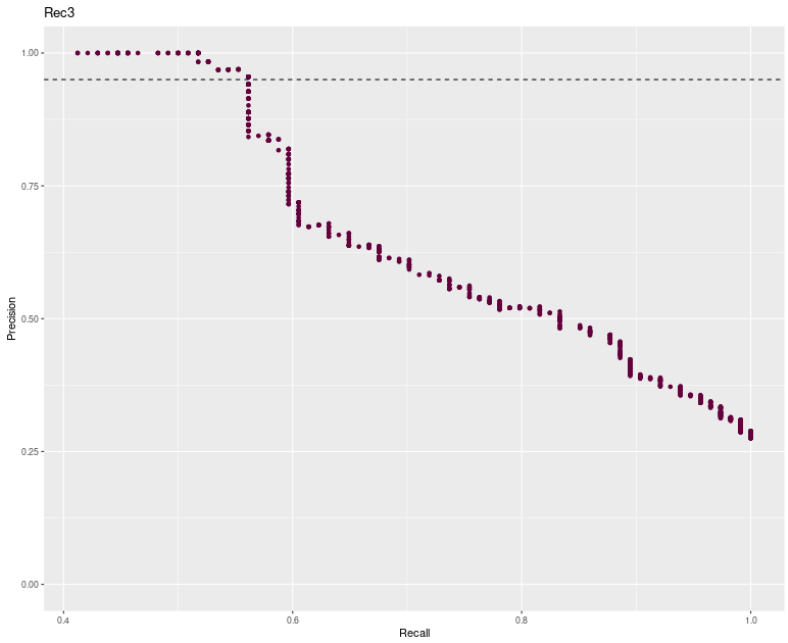


Figure 6.11 – **Performances de Colabfold-Multimer 3 recyclages pour la prédiction de PPI chez *A. thaliana* en prenant en compte les 4 meilleurs modèles**

6.2 . Discussion et perspectives

La stratégie de prédiction que j'ai mise en place, en utilisant Colabfold-Multimer avec 3 recyclages et 5 modèles permet de réduire considérablement le temps de calcul d'une prédiction sans affecter sa performance, bien au contraire. J'ai en effet montré que le fait d'utiliser 3 recyclages et plusieurs modèles permet de mieux séparer les distributions de probabilités de contact, avec des valeurs de probabilités de contact beaucoup plus faibles pour les paires aléatoires de protéines. De ce point de vue, cette stratégie améliore la prédiction. Puis, en estimant les performances de ma méthode sur un set d'un déséquilibre comparable à celui de [20], j'ai également montré que ma méthode présente une amélioration des performances par rapport à la stratégie de [20] pour une augmentation modérée du temps de calcul. Il faudrait maintenant étudier les caractéristiques de différents modèles produit par Colabfold-Multimer et notamment voir si ce sont toujours les mêmes acides aminés qui dans les différents modèles présentent la probabilité de contact maximum. C'est ce qu'on attend en effet pour les paires de protéines qui réalisent une interaction fonctionnelle. Par contre pour des protéines qui n'interagissent pas de manière fonctionnelle, il existe de nombreuses manières d'interagir et il est probable qu'aucune ne soit particulièrement privilégiée car

elle n'apporte aucune fonction à la cellule. Les acides aminés présentant la plus forte probabilité de contact devraient donc changer entre les différents modèles. Colabfold-Multimer utilisant l'information évolutive contenue dans les MSA pour construire les distogrammes desquels sont extrait la probabilité de contacts maximum, il est tout à fait possible que la variabilité des acides aminés de plus forte probabilité de contact, mesurée entre les différents modèles, soit un bon critère pour distinguer les PPI des paires aléatoires.

Dans ce chapitre, j'ai montré aussi que le fait d'utiliser Rosetta2track n'était pas si intéressant en terme de gain de temps (facteur de 1,8 entre Rosetta+colabfold et Colabfold-Multimer) par rapport aux pertes en terme de performances. Ainsi, l'exploration de l'interactome complet d'A. thaliana représente pour le moment un coût trop important en terme de temps de calcul que ce soit avec l'une ou l'autre des méthodes. Par contre, Colabfold-Multimer pourrait être utilisé pour rechercher tous les partenaires de protéines d'intérêt, impliquées dans un processus biologique donné, qu'on pourrait identifier en construisant un réseau de PPI d'ordre 2 avec APPINetwork. L'approche que j'ai mise en place permettrait par ailleurs de cibler plus précisément des paires d'intérêt en complétant et nettoyant le réseaux de PPI grâce aux probabilités d'interaction associées à chaque paires de noeuds du réseau, ce qui était un des objectifs initial de ma thèse.

Enfin, j'ai montré qu'en fixant un seuil élevé de probabilité de contact, la méthode que j'ai mise en place permet de prédire les PPI chez *A. thaliana* avec un très faible taux de faux positifs. Or c'était le second objectif de ma thèse. Tous les résultats chez *A. thaliana* ont été obtenu sur un jeu de données équilibré qui ne représente pas la réalité des données. Je ne peux donc donner aucune estimation correcte du taux de faux négatifs. Cependant sur le jeu de données équilibré, j'ai obtenu un recall de 0,56. Il y a donc au moins 44% des PPIs qui ne sont pas prédites comme telle par la méthode.

La suite de ce travail consistera donc à voir s'il est possible d'affiner les prédictions en étudiant *a posteriori* les propriétés des surfaces et des interfaces des complexes prédits afin d'identifier des caractéristiques propres au paires formant des interactions fonctionnelles. Pour cela, l'équipe dispose désormais des cartes de propriétés de surfaces pour une partie des paires de protéines des sets gold standard et aléatoires chez *S. cerevisiae*.

Enfin nous ne prédisons que des PPI de manière binaire, cependant Colabfold-Multimer permet de lancer des prédiction sur des complexes impliquant plus de protéines. Il serait intéressant d'étudier les PPI de complexes connus, en comparant les probabilités de contacts lorsqu'une paire

de protéine est prédite de manière binaire, ou lorsqu'elle est présente dans un complexe multimérique.

7 - Conclusion

L'objectif initial de ma thèse était d'étudier l'interactome d'*A. thaliana* afin d'identifier, pour des protéines de fonction inconnues, à quel processus biologique celles-ci participent. Pour cela, je devais construire cet interactome à partir des bases de données de PPI publiques et utiliser une méthode de prédiction de PPI pour le compléter et le corriger. Au cours de ma thèse, j'ai développé un outil sous forme de package R permettant de construire et d'analyser des réseaux de PPI à partir de bases de données publiques ou de données locales, APPINetwork. Cet outil m'a valu une publication en tant que premier auteur et a été mis à disposition de la communauté scientifique. J'ai ensuite travaillé sur un logiciel représentant les propriétés d'une surface protéique sur une carte en deux dimensions, ainsi que sur un pipeline permettant de construire des cartes en 2D de la propension à l'interaction de la surface d'une protéine. Ces différentes cartes devaient nous servir à prédire des PPI. Cependant, j'ai décidé de me tourner vers l'utilisation de nouvelles méthodes très prometteuses utilisant AlphaFold2 pour la prédiction de PPI. J'ai donc mis en place une méthode de prédiction de PPI en m'inspirant d'une méthode existante, mais utilisant une version plus rapide et adaptée à la prédiction de complexes, ColabFold-Multimer. J'ai étudié l'évolution des performances de cette méthode chez *S. cerevisiae* afin de trouver les paramètres les plus adaptés à la prédiction de PPI et pouvoir comparer les performances avec l'article de Humphrey et al. [20]. J'ai ainsi montré que la réduction du nombre de cycles et l'utilisation de plusieurs modèles permettaient de réduire le temps de calcul tout en améliorant les performances de la méthode de prédiction. En estimant les performances de ma méthode sur un set d'un déséquilibre comparable à celui de [20], j'ai pu montré une amélioration des performances de prédiction pour une augmentation du temps de calcul modérée. Enfin, j'ai appliqué ma méthode chez *A. thaliana* et j'ai observé des performances similaires aux prédictions chez *S. cerevisiae*. Il reste à établir un seuil sur un set plus représentatif de la réalité biologique chez *A. thaliana* en complétant le set existant avec des paires gold standard et un grand nombre de paires aléatoires avant de pouvoir réaliser des prédictions sur son interactome.

Cette thèse m'a permis de beaucoup apprendre, aussi bien sur le plan scientifique que personnel. On m'a souvent donné l'occasion de parler de mon travail au laboratoire mais aussi dans des séminaires INRAE et des colloques nationaux (JOBIM) et internationaux. J'ai ainsi progressivement pris de l'assurance. J'ai en particulier eu l'occasion de présenter mon travail aux États-Unis en présentant un poster à la conférence internationale

"Modeling of Protein Interactions". Voir l'intérêt d'autres chercheurs du domaine pour mon travail m'a permis de prendre confiance en moi sur le plan scientifique. J'ai également pu mesurer l'impact d'AlphaFold2 sur les travaux de la communauté et découvrir qu'avec le même outil, il était possible d'aborder des problématiques très différentes. Mais si j'ai pris de l'assurance, j'ai aussi appris de mes erreurs. Par exemple, si je devais refaire ma thèse aujourd'hui, je demanderais beaucoup plus tôt l'accès à des ressources de calcul nationales plutôt que de mettre en place une solution sur le cluster local de mon laboratoire. Cela m'aurait économisé beaucoup de temps et aurait permis de produire des prédictions sur un plus grand nombre de paires de protéines et des sets plus proches de la réalité biologique.

Enfin, AlphaFold2 et l'IA plus généralement, sont loin d'avoir résolu l'ensemble des problématiques liées à la prédiction des structures de complexes protéiques. Dans le cadre de la prédiction de PPI, les méthodes d'IA offrent de bonnes performances, mais il y a toujours une proportion non négligeable de PPI qui ne sont pas correctement prédites.

De plus, malgré des solutions communautaires comme ColabFold visant à populariser et accélérer ce type d'approche, les temps de calcul restent très importants si l'on veut explorer plus largement des interactomes. La méthode mise en place au cours de ma thèse semble bien plus intéressante pour compléter des réseaux de PPI à une échelle réduite, comme celle des processus biologiques, que pour l'étude d'interactome complet. Chez les plantes, il serait intéressant par exemple de prédire des PPI à l'échelle des processus de réponses aux stress. En effet, des stress tels que la sécheresse, la montée des températures ou encore le stress salin affectent la croissance des plantes. Il est donc capital d'étudier les interactions mises en jeu au cours de ces processus afin d'améliorer la réponse des plantes aux stress causés par les changements climatiques. Les méthodes d'IA et notamment de prédiction de PPI telles que ColabFold-Multimer offrent des perspectives intéressantes pour l'étude de ces processus et pourraient permettre, combinées à des réseaux de PPI, d'identifier rapidement de nouvelles protéines impliquées dans ces processus.

8 - Annexe

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim

Application Note

SURFMAP: A Software for Mapping in Two Dimensions Protein Surface Features

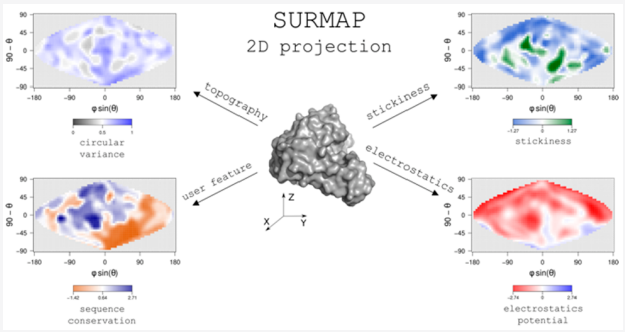
Hugo Schweke,* Marie-Hélène Mucchielli, Nicolas Chevrollier, Simon Gosset, and Anne Lopes*

Cite This: *J. Chem. Inf. Model.* 2022, 62, 1595−1601

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Molecular cartography using two-dimensional (2D) representation of protein surfaces has been shown to be very promising for protein surface analysis. Here, we present SURFMAP, a free standalone and easy-to-use software that enables the fast and automated 2D projection of either predefined features of protein surface (i.e., electrostatic potential, hydrophobicity, stickiness, and surface relief) or any descriptor encoded in the temperature factor column of a PDB file. SURFMAP proposes three different "equal-area" projections that have the advantage of preserving the area measures. It provides the user with (i) 2D maps that enable the easy and visual analysis of protein surface features of interest and (ii) maps in a text file format allowing the fast and straightforward quantitative comparison of 2D maps of homologous proteins.



INTRODUCTION

Genome sequencing has produced a huge amount of protein sequences and motivated the development of methods for protein comparison and classification. In particular, important efforts have been made to predict the function of a protein of interest from the comparison of its sequence with those of already annotated homologues. If most methods focus on information extracted from protein sequences, large-scale structural genomic experiments and homology modeling methods[1,2] have promoted the development of three-dimensional (3D) structure comparison-based methods, which are very powerful when comparing remote homologues.[3−6] Indeed, protein structure is more conserved than protein sequence, and these methods are able to highlight proteins with similar structures while sharing little sequence identity. Recent progress in protein structure prediction with AlphaFold2[7] allows the prediction of complete proteome strengthening, even more, the place of structural analyses in biology. Nevertheless, most sequence-based and structure-based methods extract information from the entire protein. It is thus difficult to distinguish residues conserved to maintain protein function from those ensuring protein stability. This is even more problematic when dealing with large families of paralogs that share the same fold but display different functions and/or physicochemical properties. Precisely, being able to predict interaction preferences (protein partners, ligands, cofactors, ...), oligomeric states or to identify thermostable proteins, for instance, would be of great value for the classification and annotation of these large protein families.

Protein surfaces can be considered as a proxy for protein interactions, thereby playing a critical role in protein function.[8] Several methods have been developed to analyze and compare protein surface features. Some of them focus on specific regions of the surface such as protein binding sites, active sites, binding pockets,[9−11,13] and ignore the rest of the surface which plays an important role in protein interactions by constantly competing with the interaction sites.[14] However, "molecular cartography" which has been introduced by Fanning et al.[15] is based on a system dimension reduction and enables the representation of the whole protein surface in a synthetic and robust way. The principle consists in projecting the 3D structure of a protein into two dimensions (2D) and then mapping features of interest (charges, hydrophobic patches, topography, sequence conservation, ...) on the resulting 2D map. Despite the efforts made on protein 3D structure representation tools,[16−18] handling 3D objects is always difficult, and these 2D maps greatly ease the visualization and analysis of the distribution of a given feature. For example, they enable the visual comparison of the distribution on homologous proteins' surfaces of specific features to understand the molecular basis of their common or different interaction properties. Additionally, 2D maps are well suited for large-scale protein surface comparisons[14] through the calculation of a map similarity with a

Received: October 17, 2021
Published: March 29, 2022

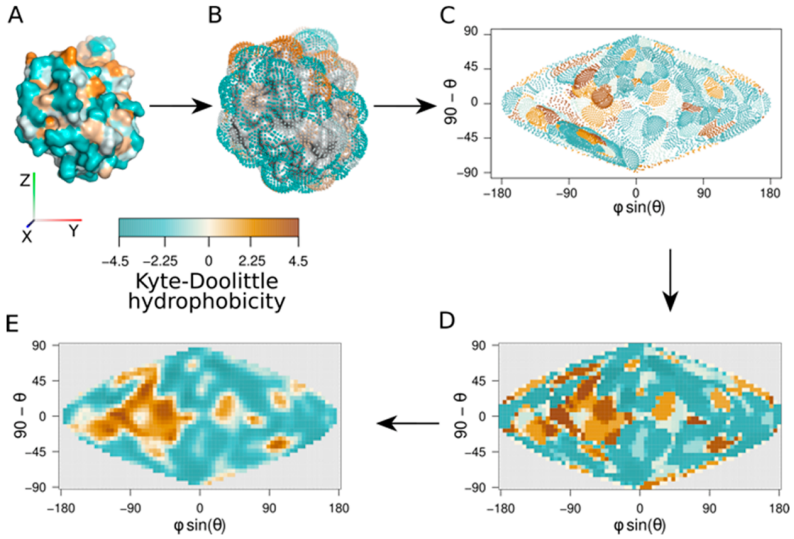


Figure 1. Generation of 2D maps. Generation of the KD hydrophobicity 2D map for chain A of the cambialistic superoxidase dismutase of *P. shermanii* (PDB code: 1ar4). (A) Each atom (or residue depending on the mapped property) is associated with its corresponding value and colored accordingly (KD hydrophobicity in this example). (B) Generation of a set of particles around the protein surface with MSMS.[32] Each particle is located at 3 Å (default) from the closest residue of the protein of interest. The KD hydrophobicity value of the closest residue of the protein is then attributed to each particle. (C) The spherical coordinates (φ, θ) (with respect to the protein center of mass) are projected onto a 2D map with the Sanson-Flamsteed 2D projection. Each projected dot is then associated with the feature value of its corresponding particle. (D) The 2D map is divided into a grid of 36 × 72 cells and each cell is associated with the average of the corresponding values. (E) Cell values are then smoothed by averaging the value of each cell with those of the eight surrounding cells. The same color scale is used for all panels (A–E).

straightforward numerical measure, and since the dimension reduction is robust against local irregularities of protein surfaces.

Although "molecular cartography" is not recent and has been shown to be very promising for protein function annotation, only a few methods have been proposed so far. Godzik et al.[8] proposed a protein surface representation based on a spherical approximation of the protein surface to compare the distribution of physicochemical features on the surface of homologous proteins. Other methods with interesting outcomes in visualization and/or surface feature comparison[19−23] are restricted to a limited subset of surface descriptors or do not provide the tool to compute the corresponding 2D maps. Recently, Kontopoulos et al., developed Structuprint,[24] a program that enables the visualization of more than 300 protein surface features. They use the Miller cylindrical projection,[25] i.e., a projection that conserves the angles and the shapes of small surfaces but that induces an important overestimation of the area at the poles. This can be problematic since the different protein surface regions do not contribute equally when comparing different surface maps, and the assignment of surface regions either to the poles or equator results from the arbitrary orientation of the protein before projection.

Here, we present SURFMAP, a program that enables the projection of predefined protein surface features: (i.e., electrostatic potential, hydrophobicity, stickiness, surface topology), residues of interest, or any descriptor encoded in the temperature factor column of a PDB file. SURFMAP enables three different equal-area 2D projections, which are either cylindrical (i.e., Lambert) or pseudocylindrical (i.e., Sanson-Flamsteed, Mollweide). These 2D projections have the advantage of preserving the area measures at the cost of distorting shapes locally. Here, we aim to preserve the size of the regions of interest rather than providing a precise representation of their shapes. Indeed, we need for a method robust against local variations since (i) proteins are flexible objects and the shapes of surface features can vary locally, (ii) protein surfaces display many local irregularities that could introduce noise when comparing two homologous surfaces, and (iii), nowadays, we frequently use 3D models that can be more or less accurate depending on the quality of the prediction. We illustrate the application of SURFMAP with different examples of analysis showing how it can be used to compare the distribution of different surface features of a protein of interest or to compare the distribution of a specific surface feature for different homologous proteins. In the latter case, the PDB files of the proteins to be compared must be prealigned before 2D projection.

METHODS

Calculation of 2D Maps. The cartography of protein surface properties is done in 5 steps. Here, we present the algorithm for the Sanson-Flamsteed 2D projection, but one should notice that SURFMAP also provides the Lambert and Mollweide 2D projections.[26]

1. Calculation of the Protein Surface Features. The first step consists in calculating the values of the feature(s) of interest for each protein residue or atom (Figure 1A). When the feature is calculated at the residue scale, each atom receives the value of the corresponding residue. Five different features can be calculated by SURFMAP. The Kyte-Doolittle (KD) and Wimley-White (WW) hydrophobicity are mapped at the residue scale and are directly derived from the Kyte-Doolittle[27] and Wimley-White[28] scales, respectively. The stickiness scale[29] reflects the propensity of each amino acid to be involved in protein−protein interfaces. The circular variance[30] (CV) can be calculated at the residue or atomic scale and measures the atom density around each atom. The CV provides a useful descriptor of the local geometry of a surface region. CV values

https://doi.org/10.1021/acs.jcim.1c01269
J. Chem. Inf. Model. 2022, 62, 1595−1601

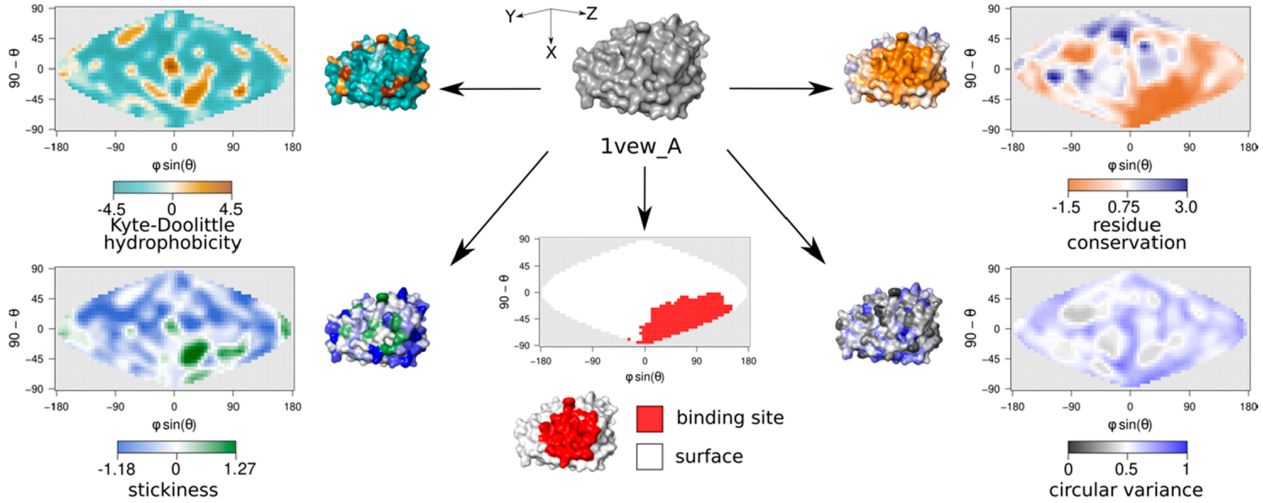


Figure 2. 2D maps of different surface features calculated for chain A of the manganese SOD of *E. coli*. The 3D structure of chain A of 1vew (PDB code: 1vew) is represented in gray (center) along with the corresponding 3D structures and Sanson-Flamsteed 2D maps colored according to the KD hydrophobicity,[27] stickiness,[29] circular variance[30] scales, and the residue conservation calculated with ConSurf[10] with conservation scores stored in the temperature column of the input PDB file. ConSurf[10] provides normalized scores where the average score of all residues of a protein is 0, the standard deviation is 1, and negative and positive scores respectively indicate positions that are more and less conserved than the average. The dimerization site is mapped with the information stored in the temperature column of the input PDB file (bottom-middle map).

range between 0 and 1 with low values reflecting protruding residues, and high values indicating residues located in cavities. The electrostatic potential of the protein surface is calculated at the atomic scale with the APBS software.[31] In addition, SURFMAP can handle any other feature stored in the temperature factor of the input PDB file.

2. Generation of a Set of Points around the Surface of the Protein. The second step consists in generating a set of particles localized at 3 Å (default value but this value can be modified by the user) from the surface of the protein of interest. This set of points is generated with MSMS,[32] a software that computes the molecular surface using a reduced surface representation of proteins (see ref 32 for more details). Here we refer to this ensemble of points as the "shell" (Figure 1B). This shell enables the precise mapping of the protein surface on a 2D map. It provides more complete information than the direct mapping of the atoms, as in the latter case the number of grid cells usually largely outnumbers the number of atoms, which results in a lot of cells left unassigned and containing no information.

3. Assignment of the Value of the Mapped Feature to Each Particle of the Shell. Each shell particle is annotated with the feature value of the closest protein atom (Figure 1B).

4. Sinusoidal Projection on a 2D Plane of the Spherical Coordinates of Each Shell Particle with Respect to the Center of Mass of the Protein. The coordinates of each shell particle p_i are expressed in spherical coordinates (φ, θ) with respect to the center of mass of the protein denoted as G (Figure S1 of the Supporting Information, SI). φ is the angle between the X axis and the projected vector of vector $(\overrightarrow{GG_i})$ in the plane $(\overrightarrow{GX}, \overrightarrow{GY})$, while θ is the angle between the vector $\overrightarrow{GG_i}$ and the Z axis. The coordinates of the shell particles are then projected onto a 2-dimensional plane by a sinusoidal projection. Each particle is therefore represented on the 2-D plane by a pair of coordinates $(x = \varphi \sin\theta, y = 90 - \theta)$, and associated with its feature value (Figure 1C) (see Recio et al.[33] for more details).

5. Division of the Map and Smoothing. The resulting map is then divided into 72 × 36 cells (the grid resolution can be modified by the user). All the 2D projections proposed by SURFMAP are equal-area. Consequently, each cell represents a surface of the same area. Each cell is then associated with the average of the particle values it contains (Figure 1D). The map can then be smoothed by averaging the score of each cell with the scores of the adjacent cells (Figure 1E).

Outputs. SURFMAP provides the user with a 2D map in a PNG or PDF format along with a text file containing the mapped values of the map. The text file can be used for numerical comparisons of 2D maps calculated (i) with different features of the same protein, or (ii) with the same feature mapped on different protein homologues.

Usage. SURFMAP is a very easy-to-use program that only needs as input the pdb file(s) of the protein(s) of interest. One should notice that if the user aims to compare surface properties of different protein homologues, their 3D structures must be prealigned to produce comparable 2D maps that are in the same frame of reference. SURFMAP enables the calculation of already encoded features such as electrostatic potential, KD or WW hydrophobicity, stickiness, CV, or can project any feature stored in the temperature factor column of the pdb file (Figure 2). SURFMAP is fast and takes on average about 3 s to generate a 2D map of a 200 residue globular protein. Electrostatic maps take a longer time (up to 30 s for a 200 residue globular protein), the limiting step being the computation of the electrostatic potential with APBS.[31]

EXAMPLE OF APPLICATIONS ON SUPEROXIDE DISMUTASES

We illustrate the application of SURFMAP with the superoxidase dismutases (SODs), a family of enzymes that catalyze the dismutation of the superoxide radical to hydrogen peroxide and molecular oxygen. Several types of SODs have been reported according to the metal ions required at the active site. Here we focus on iron (Fe), manganese (Mn), or cambialistic (Cb) SODs.[34−41] Some of the Fe and Mn-SODs form

https://doi.org/10.1021/acs.jcim.1c01269
J. Chem. Inf. Model. 2022, 62, 1595−1601

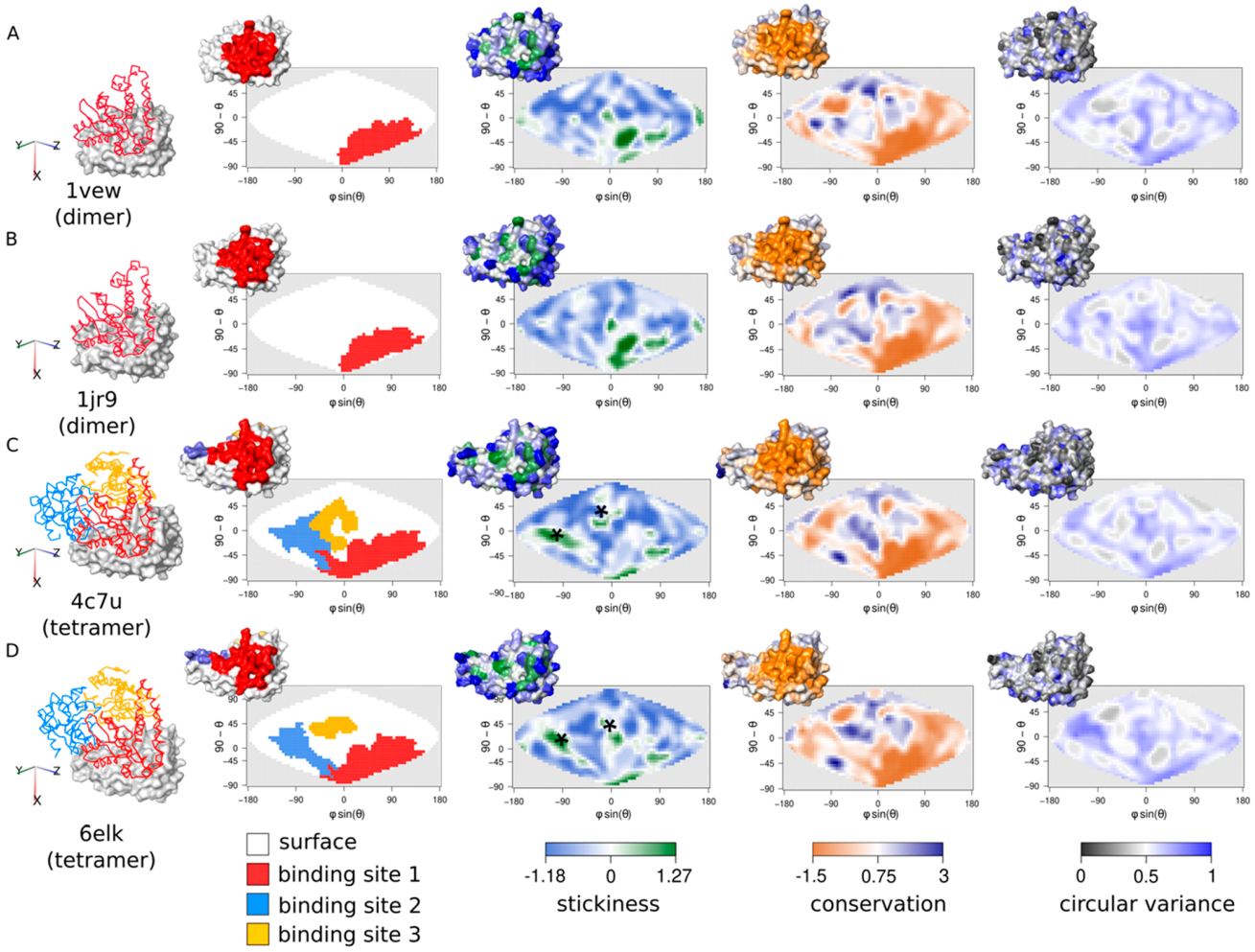


Figure 3. SOD forming homodimers display 2D maps different from those calculated for SODs forming homotetramers. Sanson-Flamsteed 2D maps calculated for four monomeric chains of Mn-SODs that either form dimers (A, B - PDB codes: 1vew and 1jr9) or homotetramers (C, D - PDB codes: 4c7u and 6elk). The 3D structure of each SOD monomer is represented with "surface" mode of pymol[18] along with the 3D structures of its corresponding interacting chains (shown with "ribbon" mode) (one interacting chain colored in red for the two dimeric SODs, and three interacting chains colored in red, blue, and yellow, respectively, for the two tetrameric SODs). For each monomer, the corresponding 2D maps representing the projection of their binding site(s) are presented (each binding site is colored according to the involved chain), along with the stickiness, residue conservation calculated with ConSurf,[10] and CV distributions over the monomer surface. Each map is also associated with the corresponding surface feature mapped on the 3D structure of the monomer. Black stars indicate sticky spots specific to the two monomers forming tetramers.

homodimers while others form homotetramers, independently of the required metal species. The experimental 3D structures of several SODs have been characterized so far and many studies have attempted to classify them to understand the molecular features explaining their Fe and Mn ion specificity, as well as their oligomeric states.[35−44] All these studies rely on the careful examination of sequence alignments or manual inspections of 3D structures.

Here, we computed the 2D maps using the Sanson-Flamsteed 2D projection of several surface descriptors of chain A of 1vew (1vew_A) which is a subunit of a dimeric Mn-SOD. Figure 2 shows the 2D projection of the interaction site participating in the dimer along with the 2D maps of KD hydrophobicity, stickiness, CV, and the sequence conservation as calculated with ConSurf.[7] ConSurf computes position-specific scores using the empirical Bayesian algorithm. These scores correspond to the evolutionary rates of each protein site and are estimated based on a multiple sequence alignment of the protein of interest with its homologues, and the corresponding tree. Conserved amino acid positions usually indicate positions of structural and/or functional importance such as positions involved in the stability of the protein or in the interaction with its partner(s). The 2D maps reveal a heterogeneous distribution of all descriptors with specific patterns distributed over the surface of the protein. The CV map is characterized by several black spots that correspond to low circular variance values and indicate protuberant regions. The stickiness reflects the interaction propensity of protein surface regions and is thereby expected to be related to hydrophobicity since protein binding sites are enriched in hydrophobic residues.

Nevertheless, if the stickiness and hydrophobicity maps are overall similar, they display specific patterns and provide complementary information (Figure 2). Indeed, the hydrophobicity map indicates a hydrophobic region (orange spot - top left) which is not predicted as sticky but rather displays an

https://doi.org/10.1021/acs.jcim.1c01269
J. Chem. Inf. Model. 2022, 62, 1595−1601

intermediate interaction propensity (white spot - top left). On the opposite, the stickiness map reveals a sticky spot (green spot on the bottom right of the map) which does not correspond to a hydrophobic region. Overall, the stickiness map displays a vast nonsticky region (top of the map and left part) along with four sticky spots that colocalize on the bottom right of the map. The latter precisely correspond to the dimer binding site (red spot on the binding site 2D projection map) which is also characterized by a flat and highly conserved region (gray region and orange spots on the circular variance and sequence conservation maps, respectively - Figure 2).

Figure 3 shows how the user can use SURFMAP to compare the surface properties of different homologous proteins. As an example, we compared several surface properties for Mn-SODs that either form homodimers or homotetramers. Therefore, we calculated their stickiness, CV, and residue conservation 2D maps along with their interaction site 2D projection maps. To produce comparable maps, all proteins were aligned beforehand in the same frame of reference with TMalign.[3] Monomers that participate in dimeric assemblies possess only one interaction site that also exists in the monomers forming tetramers (i.e., common interaction site in red). The tetramers are characterized by two additional binding sites (i.e., tetramer specific binding sites, in yellow and blue respectively), each one involving one of the three other chains of the assembly. Interestingly, if the proteins that dimerize display CV maps similar to those of proteins involved in tetramers, their stickiness and sequence conservation maps are clearly different. Indeed, the stickiness maps of the proteins forming dimers are characterized by green spots on the bottom right of the maps while the proteins forming tetramers display two additional green spots (black stars). The latter, again, correspond to the two additional binding sites involved in the tetramers (blue and yellow binding sites), supporting the stickiness descriptor as a good proxy for probing the interaction propensity of a protein surface. However, the analysis of the sequence conservation maps reveals for all monomers, a large orange spot which locates at the common binding site (red binding site), showing that its amino acids are conserved in both dimers and tetramers. That stated, the surface region that corresponds to the blue binding site specific to the tetramers is conserved to a lesser extent in monomers forming tetramers, whereas it is not the case for the corresponding nonbinding region in the proteins forming dimers. This reflects that this noninteracting region in proteins forming dimers is not subject to selection. Nevertheless, the sequences used for the estimation of the conservation of each protein, while expected to gather sequences of similar binding properties, may also contain SODs from different subfamilies (i.e., with different binding properties). One should note that the latter may blur the sequence conservation signal specific to each subfamily. Interestingly, the yellow binding site is not conserved in any of the proteins, including those able to form tetramers. To which extent this binding site is important for tetramerization deserves further investigation.

In the last section, we show an example of a quantitative comparison of 2D maps calculated for 23 SODs that form either dimers or tetramers (see PDB codes in Figure 4). Therefore, we computed the Manhattan distance between each pair of SOD 2D stickiness maps. For two maps A and B, we summed all Manhattan distances calculated between each pair of corresponding pixels a and b of coordinates (n,m) as follows:

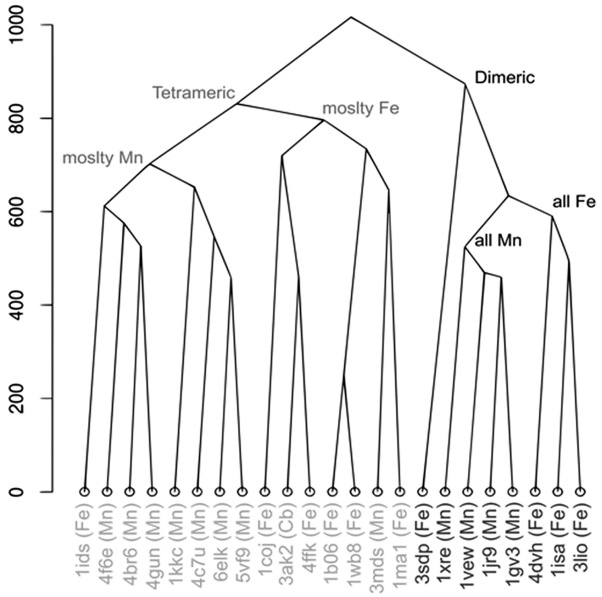


Figure 4. Distance tree based on stickiness 2D map distances. The distance tree is obtained from a hierarchical clustering algorithm (hclust method in R[45],[45] method = "complete") based on the stickiness 2D map distance matrix calculated for 23 SODs. PDB codes are indicated for each leaf along with the metal ion preference of each SOD. Monomers forming dimers are indicated in black while those forming tetramers are colored in gray. Node labels indicate the main tendency of the oligomerization state and metal binding preference of the descendant leaves.

$$d_{\mathrm{Man}}(\mathrm{A},\mathrm{B}) = \sum_{n=1}^{72}\sum_{m=1}^{36} |a_{mn} - b_{mn}| \quad (1)$$

Figure 4 shows the distance tree resulting from a Hierarchical clustering based on the map pairwise distances. Interestingly, the distance tree shows two clusters that precisely correspond to monomers forming dimers (black leaves) and tetramers respectively (gray leaves). In addition, in each oligomerization state cluster (dimers or tetramers), the SODs tend to cluster according to their metal binding preference. Interestingly, the distinction between monomers forming dimers or tetramers is less clear in the distance tree calculated from the KD hydrophobicity maps, strengthening the stickiness descriptor as a good proxy for comparing different oligomeric states between protein homologues (Figure S2).

CONCLUSIONS

We present SURFMAP, a program that enables the 2D projection of predefined protein surface descriptors or any surface feature stored in the temperature factor of a PDB file of interest. SURFMAP is a free standalone and easy-to-use software that runs on Windows, Linux, and macOS. It provides an easy and visual framework to analyze the distribution of any protein surface feature on the whole protein surface with no a priori on specific surface regions. The output 2D maps are easy to manipulate and compare, provided they are calculated in the same reference frame (i.e., the user must prealign the proteins to compare their respective maps). The 2D maps can be compared in different ways thereby enabling the user to (i) compare the distribution over the surface of a protein of interest of different features or (ii) study the evolution of a

https://doi.org/10.1021/acs.jcim.1c01269
J. Chem. Inf. Model. 2022, 62, 1595−1601

specific feature across homologous proteins. In addition, SURFMAP provides the map in a text file format allowing the fast and straightforward quantitative comparison of surface features of homologous proteins. For example, we showed with the case of SODs that comparing their 2D maps based on the stickiness feature enables the distinction of SODs with different oligomerization states and different metal ions binding preferences. In the era of protein structure prediction,[7] SURFMAP offers a powerful tool for automated and quantitative comparison of dozens of protein surfaces and opens the way to the (re)investigation and (re)annotation of large paralog families in the light of protein structure.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c01269.

Figure S1, schematic of the calculation of the spherical coordinates and Figure S2, distance tree based on KD hydrophobicity 2D map distances (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Hugo Schweke − *Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), Gif-sur-Yvette 91198, France; Department of Chemical and Structural Biology, Weizmann Institute of Science, Rehovot 7610001, Israel*; Email: hugo.schweke@weizmann.ac.il

Anne Lopes − *Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), Gif-sur-Yvette 91198, France*; orcid.org/0000-0003-0289-6608; Email: anne.lopes@i2bc.paris-saclay.fr

Authors

Marie-Hélène Mucchielli − *Université Paris-Saclay, CNRS, INRAE, Université Evry, Institute of Plant Sciences Paris-Saclay (IPS2), Gif-sur-Yvette 91190, France; Université de Paris, Institute of Plant Sciences Paris-Saclay (IPS2), Gif-sur-Yvette 91190, France*

Nicolas Chevrollier − *Independent investigator, Nyoiseau 49500, France*

Simon Gosset − *Université Paris-Saclay, CNRS, INRAE, Université Evry, Institute of Plant Sciences Paris-Saclay (IPS2), Gif-sur-Yvette 91190, France; Université de Paris, Institute of Plant Sciences Paris-Saclay (IPS2), Gif-sur-Yvette 91190, France*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c01269

Notes

The authors declare no competing financial interest.

SURFMAP can be downloaded freely at https://github.com/i2bc/SURFMAP. SURFMAP can be installed manually or through a docker container. The docker container works on Linux, macOS, and Windows, and is the recommended way to use SURFMAP. The manual installation works only on Linux distributions. SURFMAP requires MSMS,[32] and APBS[31] for electrostatics calculation. Both are provided in the docker container, but if the user chose the manual installation, the latter must be downloaded and installed separately.

■ ACKNOWLEDGMENTS

H.S. and S.G. works were supported by a French government fellowship. We thank Michel Sanner for authorizing us to integrate MSMS in a docker container. We thank Nathan Baker for authorizing us to integrate APBS in a docker container.

■ REFERENCES

(1) Guex, N.; Peitsch, M. C.; Schwede, T. Automated Comparative Protein Structure Modeling with SWISS-MODEL and Swiss-PdbViewer: A Historical Perspective. *Electrophoresis* **2009**, *30* (S1), S162−S173.

(2) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; Lepore, R.; Schwede, T. SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res.* **2018**, *46* (W1), W296−W303.

(3) Zhang, Y.; Skolnick, J. TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score. *Nucleic Acids Res.* **2005**, *33* (7), 2302−2309.

(4) Kawabata, T. MATRAS: A Program for Protein 3D Structure Comparison. *Nucleic Acids Res.* **2003**, *31* (13), 3367−3369.

(5) Malod-Dognin, N.; Pržulj, N. GR-Align: Fast and Flexible Alignment of Protein 3D Structures Using Graphlet Degree Similarity. *Bioinformatics* **2014**, *30* (9), 1259−1265.

(6) Akdel, M.; Durairaj, J.; de Ridder, D.; van Dijk, A. D. J. Caretta − A Multiple Protein Structure Alignment and Feature Extraction Suite. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 981−992.

(7) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.

(8) Pawłowski, K.; Godzik, A. Surface Map Comparison: Studying Function Diversity of Homologous Proteins. *J. Mol. Biol.* **2001**, *309* (3), 793−806.

(9) Laine, E.; Carbone, A. Local Geometry and Evolutionary Conservation of Protein Surfaces Reveal the Multiple Recognition Patches in Protein-Protein Interactions. *PLOS Comput. Biol.* **2015**, *11* (12), e1004580.

(10) Ashkenazy, H.; Abadi, S.; Martz, E.; Chay, O.; Mayrose, I.; Pupko, T.; Ben-Tal, N. ConSurf. 2016: An Improved Methodology to Estimate and Visualize Evolutionary Conservation in Macromolecules. *Nucleic Acids Res.* **2016**, *44* (W1), W344−W350.

(11) Segura, J.; Jones, P. F.; Fernandez-Fuentes, N. Improving the Prediction of Protein Binding Sites by Combining Heterogeneous Data and Voronoi Diagrams. *BMC Bioinformatics* **2011**, *12* (1), 352.

(12) Zhao, J.; Cao, Y.; Zhang, L. Exploring the Computational Methods for Protein-Ligand Binding Site Prediction. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 417−426.

(13) Xue, L. C.; Dobbs, D.; Bonvin, A. M. J. J.; Honavar, V. Computational Prediction of Protein Interfaces: A Review of Data Driven Methods. *FEBS Lett.* **2015**, *589* (23), 3516−3526.

(14) Schweke, H.; Mucchielli, M.-H.; Sacquin-Mora, S.; Bei, W.; Lopes, A. Protein Interaction Energy Landscapes Are Shaped by Functional and Also Non-Functional Partners. *J. Mol. Biol.* **2020**, *432* (4), 1183−1198.

(15) Fanning, D. W.; Smith, J. A.; Rose, G. D. Molecular Cartography of Globular Proteins with Application to Antigenic Sites. *Biopolymers* **1986**, *25* (5), 863−883.

(16) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph* **1996**, *14* (1), 33−38.

(17) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—A

Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605−1612.
(18) *The PyMOL Molecular Graphics System*, Version 1.2r3pre, Schrödinger, LLC.
(19) Pyrkov, T. V.; Chugunov, A. O.; Krylov, N. A.; Nolde, D. E.; Efremov, R. G. PLATINUM: A Web Tool for Analysis of Hydrophobic/Hydrophilic Organization of Biomolecular Complexes. *Bioinformatics* **2009**, *25* (9), 1201−1202.
(20) Koromyslova, A. D.; Chugunov, A. O.; Efremov, R. G. Deciphering Fine Molecular Details of Proteins' Structure and Function with a Protein Surface Topography (PST) Method. *J. Chem. Inf. Model* **2014**, *54* (4), 1189−1199.
(21) Levieux, G.; Tiger, G.; Mader, S.; Zagury, J.-F.; Natkin, S.; Montes, M. Udock, the Interactive Docking Entertainment System. *Faraday Discuss.* **2014**, *169* (0), 425−441.
(22) Yang, H.; Qureshi, R.; Sacan, A. Protein Surface Representation and Analysis by Dimension Reduction. *Proteome Sci.* **2012**, *10* (1), S1.
(23) Sael, L.; La, D.; Li, B.; Rustamov, R.; Kihara, D. Rapid Comparison of Properties on Protein Surface. *Proteins* **2008**, *73* (1), 1−10.
(24) Kontopoulos, D. G.; Vlachakis, D.; Tsiliki, G.; Kossida, S. Structuprint: A Scalable and Extensible Tool for Two-Dimensional Representation of Protein Surfaces. *BMC Struct. Biol.* **2016**, *16* (1), 4.
(25) Miller, O. M. Notes on Cylindrical World Map Projections. *Geogr. Rev.* **1942**, *32* (3), 424−430.
(26) Snyder, J. P. *Map Projections: A Working Manual*; Professional Paper; USGS Numbered Series 1395; U.S. Government Printing Office: Washington, D.C., 1987; Vol. *1395*. DOI: 10.3133/pp1395.
(27) Kyte, J.; Doolittle, R. F. A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.* **1982**, *157* (1), 105−132.
(28) Wimley, W. C.; White, S. H. Experimentally Determined Hydrophobicity Scale for Proteins at Membrane Interfaces. *Nat. Struct. Biol.* **1996**, *3* (10), 842−848.
(29) Levy, E. D. A Simple Definition of Structural Regions in Proteins and Its Use in Analyzing Interface Evolution. *J. Mol. Biol.* **2010**, *403* (4), 660−670.
(30) Mezei, M. A New Method for Mapping Macromolecular Topography. *J. Mol. Graph. Model* **2003**, *21* (5), 463−472.
(31) Jurrus, E.; Engel, D.; Star, K.; Monson, K.; Brandi, J.; Felberg, L. E.; Brookes, D. H.; Wilson, L.; Chen, J.; Liles, K.; Chun, M.; Li, P.; Gohara, D. W.; Dolinsky, T.; Konecny, R.; Koes, D. R.; Nielsen, J. E.; Head-Gordon, T.; Geng, W.; Krasny, R.; Wei, G.-W.; Holst, M. J.; McCammon, J. A.; Baker, N. A. Improvements to the APBS Biomolecular Solvation Software Suite. *Protein Sci. Publ. Protein Soc.* **2018**, *27* (1), 112−128.
(32) Sanner, M. F.; Olson, A. J.; Spehner, J.-C. Reduced Surface: An Efficient Way to Compute Molecular Surfaces. *Biopolymers* **1996**, *38* (3), 305−320.
(33) Fernández-Recio, J.; Totrov, M.; Abagyan, R. Identification of Protein−Protein Interaction Sites from Docking Energy Landscapes. *J. Mol. Biol.* **2004**, *335* (3), 843−865.
(34) Weatherburn, D. C. Structure and Function of Manganese-Containing Biomolecules. *Perspect. Bioinorg. Chem.* **1996**, *3*, 1−113.
(35) Cooper, J. B.; McIntyre, K.; Badasso, M. O.; Wood, S. P.; Zhang, Y.; Garbe, T. R.; Young, D. X-Ray Structure Analysis of the Iron-Dependent Superoxide Dismutase from Mycobacterium Tuberculosis at 2.0 Angstroms Resolution Reveals Novel Dimer-Dimer Interactions. *J. Mol. Biol.* **1995**, *246* (4), 531−544.
(36) Stoddard, B. L.; Howell, P. L.; Ringe, D.; Petsko, G. A. The 2.1-A Resolution Structure of Iron Superoxide Dismutase from Pseudomonas Ovalis. *Biochemistry* **1990**, *29* (38), 8885−8893.
(37) Lah, M. S.; Dixon, M. M.; Pattridge, K. A.; Stallings, W. C.; Fee, J. A.; Ludwig, M. L. Structure-Function in Escherichia Coli Iron Superoxide Dismutase: Comparisons with the Manganese Enzyme from Thermus Thermophilus. *Biochemistry* **1995**, *34* (5), 1646−1660.
(38) Ursby, T.; Adinolfi, B. S.; Al-Karadaghi, S.; De Vendittis, E.; Bocchini, V. Iron Superoxide Dismutase from the Archaeon Sulfolobus Solfataricus: Analysis of Structure and Thermostability. *J. Mol. Biol.* **1999**, *286* (1), 189−205.
(39) Edwards, R. A.; Baker, H. M.; Whittaker, M. M.; Whittaker, J. W.; Jameson, G. B.; Baker, E. N. Crystal Structure of Escherichia Coli Manganese Superoxide Dismutase at 2.1-Å Resolution. *JBIC J. Biol. Inorg. Chem.* **1998**, *3* (2), 161−171.
(40) Schmidt, M.; Meier, B.; Parak, F. X-Ray Structure of the Cambialistic Superoxide Dismutase from Propionibacterium Shermanii Active with Fe or Mn. *JBIC J. Biol. Inorg. Chem.* **1996**, *1* (6), 532−541.
(41) Knapp, S.; Kardinahl, S.; Hellgren, N.; Tibbelin, G.; Schäfer, G.; Ladenstein, R. Refined Crystal Structure of a Superoxide Dismutase from the Hyperthermophilic Archaeon Sulfolobus Acidocaldarius at 2.2 A Resolution. *J. Mol. Biol.* **1999**, *285* (2), 689−702.
(42) Hunter, T.; Bonetta, R.; Sacco, A.; Vella, M.; Sultana, P.-M.; Trinh, C. H.; Fadia, H. B. R.; Borowski, T.; Garcia-Fandiño, R.; Stockner, T.; Hunter, G. J. A Single Mutation Is Sufficient to Modify the Metal Selectivity and Specificity of a Eukaryotic Manganese Superoxide Dismutase to Encompass Iron. *Chem.-Eur. J.* **2018**, *24* (20), 5303−5308.
(43) Liao, J.; Liu, M. Y.; Chang, T.; Li, M.; Le Gall, J.; Gui, L. L.; Zhang, J. P.; Jiang, T.; Liang, D. C.; Chang, W. R. Three-Dimensional Structure of Manganese Superoxide Dismutase from Bacillus Halodenitrificans, a Component of the so-Called "Green Protein. *J. Struct. Biol.* **2002**, *139* (3), 171−180.
(44) Lim, J. H.; Yu, Y. G.; Han, Y. S.; Cho, S.; Ahn, B. Y.; Kim, S. H.; Cho, Y. The Crystal Structure of an Fe-Superoxide Dismutase from the Hyperthermophile Aquifex Pyrophilus at 1.9 A Resolution: Structural Basis for Thermostability. *J. Mol. Biol.* **1997**, *270* (2), 259−274.
(45) R Core Team: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria. **2020**.

https://doi.org/10.1021/acs.jcim.1c01269
J. Chem. Inf. Model. 2022, 62, 1595−1601

Bibliographie

[1] Minkyung Baek et al. « Accurate prediction of protein structures and interactions using a three-track neural network ». In : *Science* 373.6557 (20 août 2021), p. 871-876. issn : 0036-8075, 1095-9203. doi : 10.1126/science.abj8754. url : https://www.science.org/doi/10.1126/science.abj8754 (visité le 24/01/2023).

[2] RCSB Protein Data Bank. *PDB Statistics : Overall Growth of Released Structures Per Year*. url : https://www.rcsb.org/stats/growth/growth-released-structures (visité le 09/08/2023).

[3] Helen M. Berman et al. « The Protein Data Bank at 40 : Reflecting on the Past to Prepare for the Future ». In : *Structure* 20.3 (7 mars 2012). Publisher : Elsevier, p. 391-396. issn : 0969-2126. doi : 10.1016/j.str.2012.01.010. url : https://www.cell.com/structure/abstract/S0969-2126(12)00018-4 (visité le 09/08/2023).

[4] Nitin Bhardwaj et Hui Lu. « Correlation between gene expression profiles and protein-protein interactions within and across genomes ». In : *Bioinformatics (Oxford, England)* 21.11 (1er juin 2005), p. 2730-2738. issn : 1367-4803. doi : 10.1093/bioinformatics/bti398.

[5] Patrick Bryant, Gabriele Pozzati et Arne Elofsson. « Improved prediction of protein-protein interactions using AlphaFold2 ». In : *Nature Communications* 13.1 (10 mars 2022). Number : 1 Publisher : Nature Publishing Group, p. 1265. issn : 2041-1723. doi : 10.1038/s41467-022-28865-w. url : https://www.nature.com/articles/s41467-022-28865-w (visité le 11/01/2024).

[6] Rachel Carter et al. « Next Generation Techniques for Determination of Protein-Protein Interactions : Beyond the Crystal Structure ». In : *Current pathobiology reports* 7.3 (sept. 2019), p. 61-71. issn : 2167-485X. doi : 10.1007/s40139-019-00198-2. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7577580/ (visité le 09/08/2023).

[7] P. Chanphai, L. Bekale et H. A. Tajmir-Riahi. « Effect of hydrophobicity on protein–protein interactions ». In : *European Polymer Journal* 67 (1er juin 2015), p. 224-231. issn : 0014-3057. doi : 10.1016/j.eurpolymj.2015.03.069. url :

https://www.sciencedirect.com/science/article/pii/S001430571500213X (visité le 09/08/2023).

[8] Michael E. Cusick et al. « Literature-curated protein interaction datasets ». In : *Nature Methods* 6.1 (jan. 2009). Number : 1 Publisher : Nature Publishing Group, p. 39-46. issn : 1548-7105. doi : 10.1038/nmeth.1284. url : https://www.nature.com/articles/nmeth.1284 (visité le 20/11/2023).

[9] Marc C. Deller, Leopold Kong et Bernhard Rupp. « Protein stability : a crystallographer's perspective ». In : *Acta Crystallographica. Section F, Structural Biology Communications* 72 (Pt 2 26 jan. 2016), p. 72-95. issn : 2053-230X. doi : 10.1107/S2053230X15024619. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4741188/ (visité le 09/08/2023).

[10] Ziyun Ding et Daisuke Kihara. « Computational Methods for Predicting Protein-Protein Interactions Using Various Protein Features ». In : *Current protocols in protein science* 93.1 (août 2018), e62. issn : 1934-3655. doi : 10.1002/cpps.62. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6097941/ (visité le 22/08/2023).

[11] Richard Evans et al. *Protein complex prediction with AlphaFold-Multimer*. preprint. Bioinformatics, 4 oct. 2021. doi : 10.1101/2021.10.04.463034. url : http://biorxiv.org/lookup/doi/10.1101/2021.10.04.463034 (visité le 24/04/2023).

[12] Jane Geisler-Lee et al. « A Predicted Interactome for Arabidopsis ». In : *Plant Physiology* 145.2 (oct. 2007), p. 317-329. issn : 0032-0889. doi : 10.1104/pp.107.103465. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2048726/ (visité le 15/11/2023).

[13] Anne-Claude Gingras et al. « Analysis of protein complexes using mass spectrometry ». In : *Nature Reviews. Molecular Cell Biology* 8.8 (août 2007), p. 645-654. issn : 1471-0072. doi : 10.1038/nrm2208.

[14] Simon Gosset et al. « APPINetwork : an R package for building and computational analysis of protein–protein interaction networks ». In : *PeerJ* 10 (4 nov. 2022), e14204. issn : 2167-8359. doi : 10.7717/peerj.14204. url : https://peerj.com/articles/14204 (visité le 18/10/2023).

[15] Anna G. Green et al. « Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences ». In : *Nature Communications* 12.1 (2 mars 2021). Number : 1 Publisher : Nature Publishing Group, p. 1396. issn : 2041-1723. doi : 10.1038/s41467-021-21636-z. url : https://www.nature.com/articles/s41467-021-21636-z (visité le 19/12/2023).

[16] Yanzhi Guo et al. « Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences ». In : *Nucleic Acids Research* 36.9 (mai 2008), p. 3025-3030. issn : 0305-1048. doi : 10.1093/nar/gkn159. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2396404/ (visité le 14/11/2023).

[17] Harald Herrmann et al. « Intermediate filaments : from cell architecture to nanomechanics ». In : *Nature Reviews Molecular Cell Biology* 8.7 (juill. 2007), p. 562-573. issn : 1471-0072, 1471-0080. doi : 10.1038/nrm2197. url : https://www.nature.com/articles/nrm2197 (visité le 31/07/2023).

[18] Hailiang Huang et Joel S. Bader. « Precision and recall estimates for two-hybrid screens ». In : *Bioinformatics (Oxford, England)* 25.3 (1er fév. 2009), p. 372-378. issn : 1367-4811. doi : 10.1093/bioinformatics/btn640.

[19] Roderick E Hubbard et Muhammad Kamran Haider. « Hydrogen Bonds in Proteins : Role and Strength ». In : *Encyclopedia of Life Sciences*. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470015902.a0003011.pub2. John Wiley & Sons, Ltd, 2010. isbn : 978-0-470-01590-2. doi : 10.1002/9780470015902.a0003011.pub2. url : https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0003011.pub2 (visité le 09/08/2023).

[20] Ian R. Humphreys et al. « Computed structures of core eukaryotic protein complexes ». In : *Science* 374.6573 (10 déc. 2021), eabm4805. issn : 0036-8075, 1095-9203. doi : 10.1126/science.abm4805. url : https://www.science.org/doi/10.1126/science.abm4805 (visité le 24/04/2023).

[21] Irene M.A.Nooren et Janet M.Thornton. « Diversity of protein–protein interactions ». In : *the EMBO Journal* 22 (2003), 3486±3492. doi : 10.1093/emboj/cdg359. url : https://www.embopress.org/doi/epdf/10.1093/emboj/cdg359 (visité le 03/08/2023).

[22] John Jumper et al. « Highly accurate protein structure prediction with AlphaFold ». In : *Nature* 596.7873 (août 2021). Number : 7873 Publisher : Nature Publishing Group, p. 583-589. issn : 1476-4687. doi : 10.1038/s41586-021-03819-2. url : https://www.nature.com/articles/s41586-021-03819-2 (visité le 09/08/2023).

[23] Elizabeth Jurrus et al. « Improvements to the APBS biomolecular solvation software suite ». In : *Protein Science : A Publication of the Protein Society* 27.1 (jan. 2018), p. 112-128. issn : 1469-896X. doi : 10.1002/pro.3280.

[24] Minoru Kanehisa et al. « KEGG as a reference resource for gene and protein annotation ». In : *Nucleic Acids Research* 44 (D1 4 jan. 2016), p. D457-D462. issn : 0305-1048. doi : 10.1093/nar/gkv1070. url : https://doi.org/10.1093/nar/gkv1070 (visité le 14/11/2023).

[25] J. C. Kendrew et al. « A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis ». In : *Nature* 181.4610 (mars 1958). Number : 4610 Publisher : Nature Publishing Group, p. 662-666. issn : 1476-4687. doi : 10.1038/181662a0. url : https://www.nature.com/articles/181662a0 (visité le 09/08/2023).

[26] Dima Kozakov et al. « The ClusPro web server for protein-protein docking ». In : *Nature protocols* 12.2 (fév. 2017), p. 255-278. issn : 1754-2189. doi : 10.1038/nprot.2016.169. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5540229/ (visité le 08/01/2024).

[27] Andriy Kryshtafovych et al. « Critical Assessment of Methods of Protein Structure Prediction (CASP) – Round XIII ». In : *Proteins* 87.12 (déc. 2019), p. 1011-1020. issn : 0887-3585. doi : 10.1002/prot.25823. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6927249/ (visité le 19/10/2023).

[28] J. Kyte et R. F. Doolittle. « A simple method for displaying the hydropathic character of a protein ». In : *Journal of Molecular Biology* 157.1 (5 mai 1982), p. 105-132. issn : 0022-2836. doi : 10.1016/0022-2836(82)90515-0.

[29] P. Therese Lang et al. « DOCK 6 : combining techniques to model RNA-small molecule complexes ». In : *RNA (New York, N.Y.)* 15.6 (juin 2009), p. 1219-1230. issn : 1469-9001. doi : 10.1261/rna.1563609.

[30] Guillaume Launay, Nicoletta Ceres et Juliette Martin. « Non-interacting proteins may resemble interacting proteins : prevalence and implications ». In : *Scientific Reports* 7.1 (13 jan. 2017). Number : 1 Publisher : Nature Publishing Group, p. 40419. issn : 2045-2322. doi : 10.1038/srep40419. url : https://www.nature.com/articles/srep40419 (visité le 01/08/2023).

[31] Emmanuel D. Levy. « A simple definition of structural regions in proteins and its use in analyzing interface evolution ». In : *Journal of Molecular Biology* 403.4 (5 nov. 2010), p. 660-670. issn : 1089-8638. doi : 10.1016/j.jmb.2010.09.028.

[32] Jer-Sheng Lin et Erh-Min Lai. « Protein–Protein Interactions : Co-Immunoprecipitation ». In : *Bacterial Protein Secretion Systems*. Sous la dir. de Laure Journet et Eric Cascales. T. 1615. Series Title : Methods in Molecular Biology. New York, NY : Springer New York, 2017, p. 211-219. isbn : 978-1-4939-7031-5 978-1-4939-7033-9. doi : 10.1007/978-1-4939-7033-9_17. url : http://link.springer.com/10.1007/978-1-4939-7033-9_17 (visité le 07/08/2023).

[33] Mingzhi Lin, Xueling Shen et Xin Chen. « PAIR : the predicted Arabidopsis interactome resource ». In : *Nucleic Acids Research* 39 (Database issue jan. 2011), p. D1134-D1140. issn : 0305-1048. doi : 10.1093/nar/gkq938. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013789/ (visité le 18/08/2023).

[34] Marcos Malumbres. « Cyclin-dependent kinases ». In : *Genome Biology* 15.6 (2014), p. 122. issn : 1465-6906. doi : 10.1186/gb4184. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4097832/ (visité le 27/07/2023).

[35] Valerio Mariani et al. « lDDT : a local superposition-free score for comparing protein structures and models using distance difference tests ». In : *Bioinformatics* 29.21 (1er nov. 2013), p. 2722-2728. issn : 1367-4803. doi : 10.1093/bioinformatics/btt473. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3799472/ (visité le 26/10/2023).

[36] T. W. Martin et Zygmunt S. Derewenda. « The name is bond — H bond ». In : *Nature Structural Biology* 6.5 (mai 1999). Number : 5 Publisher : Nature Publishing Group, p. 403-406. issn : 1545-9985. doi : 10.1038/8195. url : https://www.nature.com/articles/nsb0599_403 (visité le 09/08/2023).

[37] Elisa Martino et al. « Mapping, Structure and Modulation of PPI ». In : *Frontiers in Chemistry* 9 (2021). issn : 2296-2646. url : https://www.frontiersin.org/articles/10.3389/fchem.2021.718405 (visité le 07/08/2023).

[38] Richard H. McCoy, Curtis E. Meyer et William C. Rose. « FEEDING EXPERIMENTS WITH MIXTURES OF HIGHLY PURIFIED AMINO ACIDS ». In : *Journal of Biological Chemistry* 112.1 (déc. 1935), p. 283-302. issn : 00219258. doi : 10.1016/S0021-9258(18)74986-7. url : https://linkinghub.elsevier.com/retrieve/pii/S0021925818749867 (visité le 09/08/2023).

[39] Mihaly Mezei. « A new method for mapping macromolecular topography ». In : *Journal of Molecular Graphics & Modelling* 21.5 (mars 2003), p. 463-472. issn : 1093-3263. doi : 10.1016/s1093-3263(02)00203-6.

[40] Milot Mirdita et al. « ColabFold : making protein folding accessible to all ». In : *Nature Methods* 19.6 (juin 2022), p. 679-682. issn : 1548-7091, 1548-7105. doi : 10.1038/s41592-022-01488-1. url : https://www.nature.com/articles/s41592-022-01488-1 (visité le 22/05/2023).

[41] Isabel Moraes et al. « Membrane protein structure determination — The next generation ». In : *Biochimica et Biophysica Acta* 1838.1 (jan. 2014), p. 78-87. issn : 0006-3002. doi : 10.1016/j.bbamem.2013.07.010. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3898769/ (visité le 09/08/2023).

[42] Roberto Mosca, Arnaud Céol et Patrick Aloy. « Interactome3D : adding structural details to protein networks ». In : *Nature Methods* 10.1 (jan. 2013), p. 47-53. issn : 1548-7105. doi : 10.1038/nmeth.2289.

[43] G. J. Mulder. « On the composition of some animal substances ». In : *Journal fur Praktische Chemie* 16.1 (1839), p. 129-152. issn : 0021-8383, 1521-3897. doi : 10.1002/prac.18390160137. url : https://onlinelibrary.wiley.com/doi/10.1002/prac.18390160137 (visité le 17/07/2023).

[44] Vic E. Myer et Richard A. Young. « RNA Polymerase II Holoenzymes and Subcomplexes * ». In : *Journal of Biological Chemistry* 273.43 (23 oct. 1998). Publisher : Elsevier, p. 27757-27760. issn : 0021-9258, 1083-351X. doi : 10.1074/jbc.273.43.27757. url : https://www.jbc.org/article/S0021-9258(19)59524-2/abstract (visité le 31/07/2023).

[45] Lydia Nisius et Stephan Grzesiek. « Key stabilizing elements of protein structure identified through pressure and temperature perturbation of its hydrogen bond network ». In : *Nature Chemistry* 4.9 (sept. 2012), p. 711-717. issn : 1755-4330, 1755-4349. doi : 10.1038/nchem.1396. url : https://www.nature.com/articles/nchem.1396 (visité le 09/08/2023).

[46] Sandra Orchard et al. « Protein Interaction Data Curation - The International Molecular Exchange Consortium (IMEx) ». In : *Nature methods* 9.4 (avr. 2012), p. 345-350. issn : 1548-7091. doi : 10.1038/nmeth.1931. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3703241/ (visité le 09/10/2020).

[47] Sandra Orchard et al. « The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases ». In : *Nucleic Acids Research* 42 (D1 jan. 2014), p. D358-D363. issn : 0305-1048, 1362-4962. doi : 10.1093/nar/gkt1115. url : https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1115 (visité le 12/10/2020).

[48] Rose Oughtred et al. « The BioGRID interaction database : 2019 update ». In : *Nucleic Acids Research* 47 (D1 8 jan. 2019), p. D529-D541. issn : 0305-1048, 1362-4962. doi : 10.1093/nar/gky1079. url : https://academic.oup.com/nar/article/47/D1/D529/5204333 (visité le 12/10/2020).

[49] Typhaine Paysan-Lafosse et al. « InterPro in 2022 ». In : *Nucleic Acids Research* 51 (D1 6 jan. 2023), p. D418-D427. issn : 0305-1048. doi : 10.1093/nar/gkac993. url : https://doi.org/10.1093/nar/gkac993 (visité le 14/11/2023).

[50] F. Pazos et A. Valencia. « Similarity of phylogenetic trees as indicator of protein-protein interaction ». In : *Protein Engineering* 14.9 (sept. 2001), p. 609-614. issn : 0269-2139. doi : 10.1093/protein/14.9.609.

[51] Philippe Gambette et Alain Guénoche. « Bootstrap clustering for graph partitioning ». In : *RAIRO - Operations Research* 45.4 (oct. 2011), p. 339-352. issn : 0399-0559, 1290-3868. doi : 10.1051/ro/2012001. url : http://www.rairo-ro.org/10.1051/ro/2012001 (visité le 19/08/2020).

[52] Philippe Lamesch et al. « The Arabidopsis Information Resource (TAIR) : improved gene annotation and new tools ». In : *Nucleic Acids Research* 40 (Database issue jan. 2012), p. D1202-D1210. issn : 0305-1048. doi : 10.1093/nar/gkr1090. url : https:

//www.ncbi.nlm.nih.gov/pmc/articles/PMC3245047/ (visité le 20/08/2020).

[53] Brian G. Pierce et al. « ZDOCK server : interactive docking prediction of protein–protein complexes and symmetric multimers ». In : *Bioinformatics* 30.12 (15 juin 2014), p. 1771-1773. issn : 1367-4803. doi : 10.1093/bioinformatics/btu097. url : https://doi.org/10.1093/bioinformatics/btu097 (visité le 08/01/2024).

[54] Sylvain Pitre et al. « PIPE : a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs ». In : *BMC bioinformatics* 7 (27 juill. 2006), p. 365. issn : 1471-2105. doi : 10.1186/1471-2105-7-365.

[55] Sabry Razick, George Magklaras et Ian M Donaldson. « iRefIndex : A consolidated protein interaction database with provenance ». In : *BMC Bioinformatics* 9 (30 sept. 2008), p. 405. issn : 1471-2105. doi : 10.1186/1471-2105-9-405. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2573892/ (visité le 08/11/2023).

[56] C.M. Roth, B.L. Neal et A.M. Lenhoff. « Van der Waals interactions involving proteins ». In : *Biophysical Journal* 70.2 (fév. 1996), p. 977-987. issn : 00063495. doi : 10.1016/S0006-3495(96)79641-8. url : https://linkinghub.elsevier.com/retrieve/pii/S0006349596796418 (visité le 09/08/2023).

[57] F. Sanger. « The terminal peptides of insulin ». In : *Biochemical Journal* 45.5 (1949), p. 563-574. issn : 0264-6021. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1275055/ (visité le 09/08/2023).

[58] M. F. Sanner, A. J. Olson et J. C. Spehner. « Reduced surface : an efficient way to compute molecular surfaces ». In : *Biopolymers* 38.3 (mars 1996), p. 305-320. issn : 0006-3525. doi : 10.1002/(SICI)1097-0282(199603)38:3%3C305::AID-BIP4%3E3.0.CO;2-Y.

[59] Hugo Schweke. « Développement d'une méthode in silico pour caractériser le potentiel d'interaction des surfaces protéiques dans un environnement encombré ». In : *Université Paris-Sud* (2018).

[60] Hugo Schweke et al. « SURFMAP : A Software for Mapping in Two Dimensions Protein Surface Features ». In : *Journal of Chemical Information and Modeling* 62.7 (11 avr. 2022). Publisher : American Chemical Society, p. 1595-1601. issn : 1549-9596. doi : 10.1021/acs.jcim.1c01269. url : https://doi.org/10.1021/acs.jcim.1c01269 (visité le 18/10/2023).

[61] Andrew W. Senior et al. « Improved protein structure prediction using potentials from deep learning ». In : *Nature* 577.7792 (jan. 2020). Number : 7792 Publisher : Nature Publishing Group, p. 706-710. issn : 1476-4687. doi : 10.1038/s41586-019-1923-7. url : https://www.nature.com/articles/s41586-019-1923-7 (visité le 19/10/2023).

[62] I. G. Serebriiskii et E. A. Golemis. « Two-hybrid system and false positives. Approaches to detection and elimination ». In : *Methods in Molecular Biology (Clifton, N.J.)* 177 (2001), p. 123-134. issn : 1064-3745. doi : 10.1385/1-59259-210-4:123.

[63] Christian J. A. Sigrist et al. « New and continuing developments at PROSITE ». In : *Nucleic Acids Research* 41 (D1 1er jan. 2013), p. D344-D347. issn : 0305-1048. doi : 10.1093/nar/gks1067. url : https://doi.org/10.1093/nar/gks1067 (visité le 14/11/2023).

[64] Ta-tsen Soong, Kazimierz O. Wrzeszczynski et Burkhard Rost. « Physical protein–protein interactions predicted from microarrays ». In : *Bioinformatics* 24.22 (15 nov. 2008), p. 2608-2614. issn : 1367-4803. doi : 10.1093/bioinformatics/btn498. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2579715/ (visité le 15/11/2023).

[65] Damian Szklarczyk et al. « STRING v11 : protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets ». In : *Nucleic Acids Research* 47 (Database issue 8 jan. 2019), p. D607-D613. issn : 0305-1048. doi : 10.1093/nar/gky1131. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6323986/ (visité le 09/10/2020).

[66] Gohar Taj et al. « MAPK machinery in plants ». In : *Plant Signaling & Behavior* 5.11 (nov. 2010), p. 1370-1378. issn : 1559-2316. doi : 10.4161/psb.5.11.13020. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3115236/ (visité le 27/07/2023).

[67] The UniProt Consortium. « UniProt : the Universal Protein Knowledgebase in 2023 ». In : *Nucleic Acids Research* 51 (D1 6 jan. 2023), p. D523-D531. issn : 0305-1048. doi : 10.1093/nar/gkac1052. url : https://doi.org/10.1093/nar/gkac1052 (visité le 14/11/2023).

[68] Alba Timón-Gómez et al. « Mitochondrial Cytochrome c Oxidase Biogenesis : Recent Developments ». In : *Seminars in cell & developmental biology* 76 (avr. 2018), p. 163-178. issn : 1084-9521. doi : 10.1016/j.semcdb.2017.08.055. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5842095/ (visité le 19/11/2023).

[69] Ashutosh Tripathi et Vytas A Bankaitis. « Molecular Docking : From Lock and Key to Combination Lock ». In : *Journal of molecular medicine and clinical applications* 2.1 (2017), p. 10.16966/2575-0305.106. issn : 2575-0305. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5764188/ (visité le 09/08/2023).

[70] Nurcan Tuncbag et al. « Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM ». In : *Nature protocols* 6.9 (11 août 2011), p. 1341-1354. issn : 1754-2189. doi : 10.1038/nprot.2011.367. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7384353/ (visité le 15/11/2023).

[71] Filippo Vascon et al. « Protein electrostatics : From computational and structural analysis to discovery of functional fingerprints and biotechnological design ». In : *Computational and Structural Biotechnology Journal* 18 (1er jan. 2020), p. 1774-1789. issn : 2001-0370. doi : 10.1016/j.csbj.2020.06.029. url : https://www.sciencedirect.com/science/article/pii/S2001037020303184 (visité le 08/08/2023).

[72] Ashish Vaswani et al. « Attention is All you Need ». In : ().

[73] M. Vauquelin. « XLV. Discovery of a new vegetable principle in asparagus (Asparagus sativus of Linnæus) ». In : (1er jan. 1807). doi : 10.1080/14786440708563676. url : https://zenodo.org/record/1660477 (visité le 09/08/2023).

[74] Sjoerd de Vries et Martin Zacharias. « Flexible docking and refinement with a coarse-grained protein model using ATTRACT : Flexible Protein-Protein Docking and Refinement ». In : *Proteins : Structure, Function, and Bioinformatics* 81.12 (déc. 2013), p. 2167-2174. issn : 08873585. doi : 10.1002/prot.24400.

url : https://onlinelibrary.wiley.com/doi/10.1002/prot.24400 (visité le 24/04/2023).

[75] Mark Nicholas Wass et al. « Towards the prediction of protein interaction partners using physical docking ». In : *Molecular Systems Biology* 7 (15 fév. 2011), p. 469. issn : 1744-4292. doi : 10.1038/msb.2011.3. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3063693/ (visité le 15/11/2023).

[76] W. C. Wimley et S. H. White. « Experimentally determined hydrophobicity scale for proteins at membrane interfaces ». In : *Nature Structural Biology* 3.10 (oct. 1996), p. 842-848. issn : 1072-8368. doi : 10.1038/nsb1096-842.

[77] Haiyuan Yu et al. « Predicting interactions in protein networks by completing defective cliques ». In : *Bioinformatics (Oxford, England)* 22.7 (1er avr. 2006), p. 823-829. issn : 1367-4803. doi : 10.1093/bioinformatics/btl014.

[78] Evgeny M Zdobnov et al. « OrthoDB in 2020 : evolutionary and functional annotations of orthologs ». In : *Nucleic Acids Research* 49 (D1 8 jan. 2021), p. D389-D393. issn : 0305-1048. doi : 10.1093/nar/gkaa1009. url : https://doi.org/10.1093/nar/gkaa1009 (visité le 14/11/2023).

[79] Qiangfeng Cliff Zhang et al. « Structure-based prediction of protein–protein interactions on a genome-wide scale ». In : *Nature* 490.7421 (25 oct. 2012), p. 556-560. issn : 0028-0836, 1476-4687. doi : 10.1038/nature11503. url : http://www.nature.com/articles/nature11503 (visité le 02/10/2020).

[80] Shu-Bo Zhang et Qiang-Rong Tang. « Protein–protein interaction inference based on semantic similarity of Gene Ontology terms ». In : *Journal of Theoretical Biology* 401 (21 juill. 2016), p. 30-37. issn : 0022-5193. doi : 10.1016/j.jtbi.2016.04.020. url : https://www.sciencedirect.com/science/article/pii/S0022519316300480 (visité le 15/11/2023).

[81] Yang Zhang et Jeffrey Skolnick. « TM-align : a protein structure alignment algorithm based on the TM-score ». In : *Nucleic Acids Research* 33.7 (2005), p. 2302-2309. issn : 0305-1048. doi : 10.1093/nar/gki524. url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1084323/ (visité le 26/10/2023).

[82] Yanyan Zhao et al. « CryoEM reveals the stochastic nature of individual ATP binding events in a group II chaperonin ». In : *Nature Communications* 12.1 (6 août 2021), p. 4754. issn : 2041-1723. doi : 10.1038/s41467-021-25099-0.

[83] Yuan Zhou et al. « Can simple codon pair usage predict protein-protein interaction? » In : *Molecular bioSystems* 8.5 (avr. 2012), p. 1396-1404. issn : 1742-2051. doi : 10.1039/c2mb05427b.

[84] G. C. P. van Zundert et al. « The HADDOCK2.2 Web Server : User-Friendly Integrative Modeling of Biomolecular Complexes ». In : *Journal of Molecular Biology*. Computation Resources for Molecular Biology 428.4 (22 fév. 2016), p. 720-725. issn : 0022-2836. doi : 10.1016/j.jmb.2015.09.014. url : https://www.sciencedirect.com/science/article/pii/S0022283615005379 (visité le 05/01/2024).