UNIVERSITE PARIS-SUD

Ecole Doctorale de Physique de la Règion Parisienne – ED 107

laboratoire de Physique Thèorique et Modèles Statistiques

DISCIPLINE: Physique

SYTHÈSE

par

Olga VALBA

Statistical analysis of networks and biophysical systems of complex architecture

---

L'analyse statistique des réseaux et des systmes biophysiques de l'architecture complexe

---

1 Introduction

De nombreux systèmes biologiques présentent une organisation complexe. Par exemple, les biopolymères peuvent posséder une structure très hiérarchisée responsable de leur fonction particulière. Comprendre la complexité de cette organisation permet de décrire des phénomènes biologiques et de prédire les fonctions des molécules. En outre, en supposant que la structure primaire du polymère est formée aléatoirement, nous pouvons essayer de caractériser ce phénomène par des grandeurs probabilistes (variances, moyennes, etc). Cette formulation est propre aux problèmes d'évolution. Les réseaux biologiques sont d'autres objets communs de la physique statistique possédant de riches propriétés fonctionnelles. Pour décrire un mécanisme biologique, on utilise différents types de réseaux biomoléculaires. Le développement de nouvelles approches peut nous aider à structurer, représenter et interpréter des données expérimentales, comprendre les processus cellulaires et prédire la fonction d'une molécule.

L'objectif de cette thèse est de développer des méthodes pour l'étude d'objets statiques ou dynamiques, ayant une architecture complexe. Ici, nous nous intéressons à deux problèmes.

La première partie est consacrée à l'analyse statistique des biopolymères aléatoires. Nous étudions une transition de phase présente dans les séquences aléatoires de l'ARN. On met alors en évidence deux modes: le régime où presque toutes les bases qui composent l'ARN sont couplées et la situation où une fraction finie de ces bases restent non complémentaires.

La deuxième partie de cette thèse se concentre sur les propriétés statistiques des réseaux. Nous développons des méthodes pour l'identification d'amas de gènes co-expressifs sur les réseaux et la prédiction de gènes régulateurs novateurs. Pour cela, nous utilisons la fonction du plus court chemin et l'analyse du profil des motifs formés par ces amas. Ces méthodes ont pu prédire les facteurs de transcription impliqués dans le processus de longévité. Enfin, nous discutons de la formation de motifs stables sur les réseaux due à une évolution sélective.

2 Polymères aléatoires

2.1 Description du modèle

Les molécules de l'ARN forment une structure secondaire complexe. Les particularités principales que nous utilisons dans l'analyse sont les suivantes:

1. Les liaisons entre les monomères dans la structure secondaire se constituent selon les règles de complémentarité

2. La structure secondaire de l'ARN a une structure encapsulée hiérarchique (Fig. 1). Des pseudonuds sont rares, donc nous ignorons leur discussion.

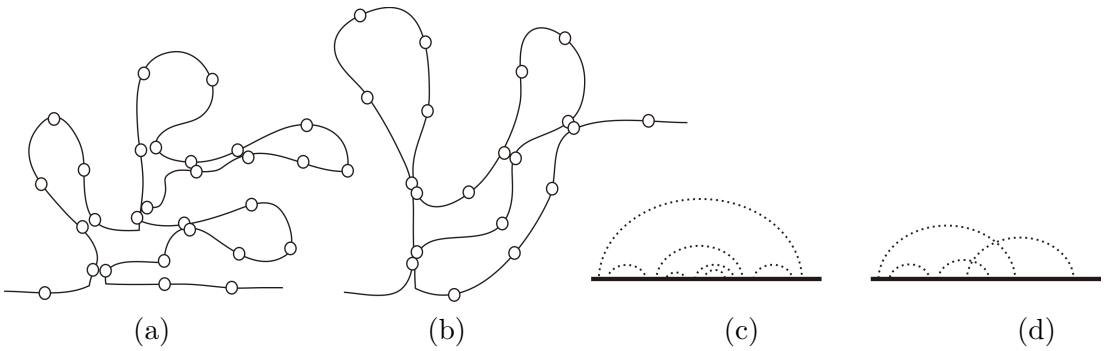


Figure 1. Structure de trèfle de l'ARN (a) et pseudonud (b); (c) et (d) son les représentations en arc respectivement.

Analysons le modèle statistique auxiliaire décrivant la formation de la structure secondaire du polymère avec une séquence primaire arbitraire. On pose la longueur de ce polymère L, mesurée en unités de charnières monomères. Chaque monomère peut être choisi de c monomères différents A,B,C,D,... (Pour les séquences de l'ARN, c = 4).

Aux températures suffisamment basses où la contribution entropique peut être ignorée en comparaison de la contribution énergétique, la somme statistique de la chane peut être présentée comme suit

$$G_{1,L} = 1 + \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} e^{\epsilon_{i,j}/T} G_{i+1,j-1}\, G_{j+1,n}$$

Où $\epsilon_{i,j}$ décrit l'énergie d'interaction entre les monomères i et j. Sans perdre de communalité, nous estimons $\epsilon_{i,j} = 1$ pour la paire complémentaire de bases et $\epsilon_{i,j} = 0$ en cas de liaison non-complémentaire. A son tour, la somme statistique

est liée avec l'énergie libre du complexe $G_{1,L}$ et avec la température T au moyen de la relation connue $G_{1,L} = \exp\{F_{1,L}/T\}$. En passant dans l'équation récursive à la somme statistique vers la limite $T \to 0$ [1], on obtient

$$F_{i,i+k} = \max_{s} \left[F_{i+1,i+k}, (F_{i+1,s-1} + F_{s+1,i+k} + \epsilon_{i,s})\right]$$

Cette transition est bien légitime, vu que l'énergie de la paire complémentaire dépasse la température ambiante de dizaines de fois. Cette équation de programmation dynamique décrit la topologie requise encapsulée de la structure secondaire de l'ARN [2].

2.2 Propriétés statistiques

Des séquences primaires aléatoires de l'ARN à alphabet varié c ont fait l'objet principal de notre analyse. Nous montrons quil existe une conduite critique dans le système du polymère aléatoire pareil à l'ARN en fonction de l'alphabet qui y est utilisé [3]. Le point critique divise les deux modes: le premier mode est caractérisé par une structure secondaire idéale sans lacunes (Fig. 2(b)), alors que la phase post-critique a toujours une part finale de monomères libres (Fig. 2(a)). Les chemins sans portions horizontales (chemins de Dyck) correspondent aux premières structures et les dits chemins de Motzkin correspondent aux secondes [4].

Les estimations analytiques basées sur la comparaison des ensembles de toutes les structures idéales (Fig. 2) et de toutes les séquences aléatoires de l'alphabet c ont montré que la valeur critique de l'alphabet se trouvait dans la plage $2 < c_c < 4$. Pour déterminer l'alphabet critique, le modèle du polymère Bernoulli a été proposé, où la matrice de contacts possibles $\epsilon = V$ est aléatoire avec la probabilité p. Cet examen permet de générer un polymère aléatoire avec toute valeur effective de l'alphabet. Une expérience numérique (Fig. 3) montre que les valeurs d'énergie limite f_∞ pour l'alphabet Bernoulli ne sont pas différents de plus de 1% des valeurs correspondantes pour les séquences alphabétiques (avec la valeur discrète de l'alphabet c), ce qui justifie l'utilisation de ce modèle.

Pour déterminer numériquement le point de transition, nous avons analysé les ensembles de polymères aléatoires ($N = 10^4$ des séquences primaires dans chaque ensemble) de différentes longueurs, et nous avons réalisé un scaling des courbes

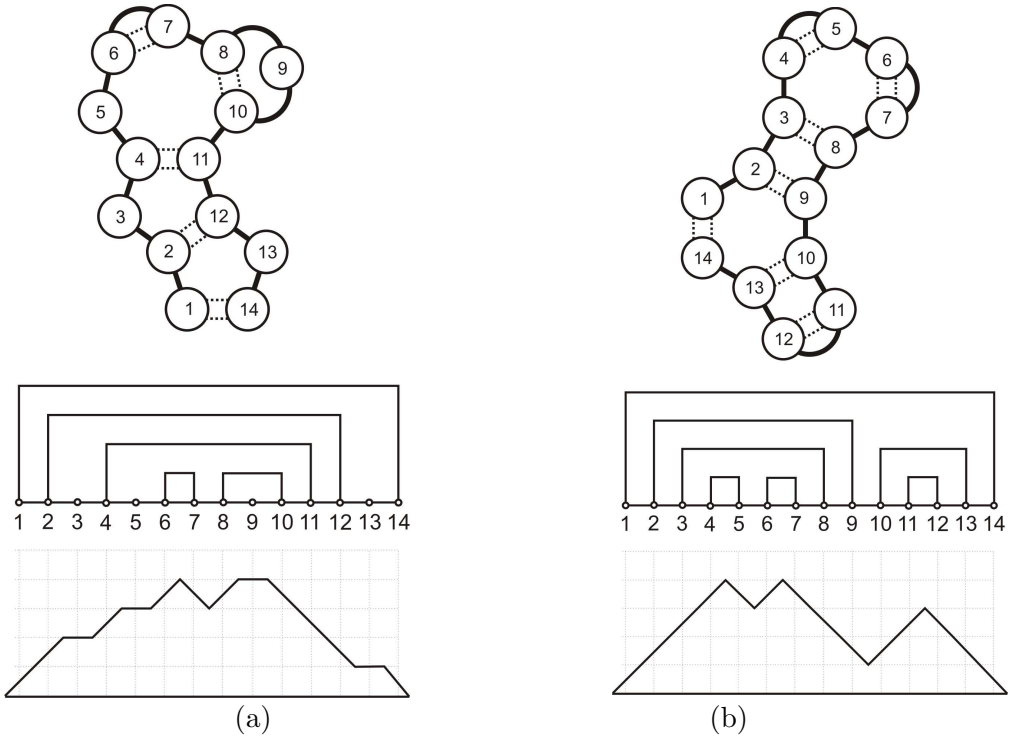


Figure 2. Structure secondaire de l'ARN avec lacunes (a) et sans lacunes (b) et cheminement aléatoire qui leur correspond.

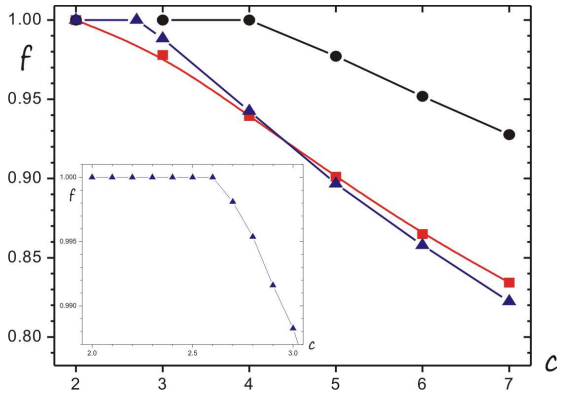


Figure 3. Dépendance de la valeur limite d'énergie de l'alphabet pour séquences alphabétiques (rouge) et pour le polymère aléatoire Bernoulli (bleu), et estimation analytique de l'énergie en haut dans le modèle de cheminement aléatoire.

résultantes (Fig. 4 (b)). En résultat de cette procédure, on a reu la valeur de l'alphabet critique $c_c = 2.67$.

Les phases sous-critique et post-critique sont caractérisées par les propriétés statistiques différentes, notamment, par le rapprochement différent de l'énergie

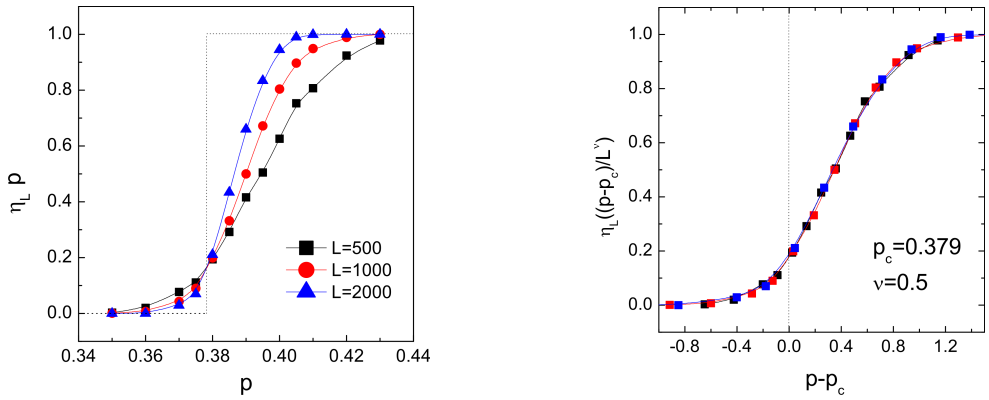


Figure 4. Dépendance de la probabilité de la structure idéale dans l'ensemble de structures de différentes longueurs (a) et analyse scaling des dépendances obtenues (b).

spécifique f_∞ de sa valeur limite:

$$\begin{cases} f_\infty(p) - f_L(p) \sim e^{-L/\ell(p)} & \text{pour } p > p_c \\ f_\infty(p) - f_L(p) \sim L^{-\alpha(p)} & \text{pour } p < p_c \end{cases}$$

Le modèle Bernoulli permet de réaliser une estimation analytique plus précise du point de transition [5]. Pour l'estimation, nous admettons que la configuration typique d'arc se constitue comme suit:

1. Sélection de $L/4$ arcs courts disjoints de $L-1$ arcs possibles selon la matrice de contacts V

2. Sélection indépendante de $L/4$ arcs restants des arcs longs

Cette procédure est due au fait que des arcs de différentes longueurs se rencontrent dans la configuration planaire optimale avec probabilité variée, en particulier, la probabilité du plus court arc $P(i, i+1) = \frac{1}{4}$. Le dégagement des plus courts arcs dans la structure idéale est tenu en compte immédiatement par le calcul de la probabilité de sélection de $L/4$ arcs des pL possibles (admettant que les arcs permis sont distribués régulièrement le long de la chane). Cette estimation combinatoire mène à l'estimation $c_c = 2.87$, ce qui est proche du nombre observé.

l'ouvrage montre aussi une corrélation entre la transition de phase thermodynamique et la transition structurelle étudiée. Comme on a déjà vu [6], en fonction de la température, l'ARN aléatoire se trouve dans une des phases: i) phase fondue (de haute température) ou ii) phase glaciale (de basse température). Dans

la phase de haute température, l'entropie de la chane a un rle plus important que l'ordre des monomères dans la structure primaire qui détermine l'énergie de la séquence. La phase de basse température est déterminée par un désordre dans la séquence primaire. Il a été démontré que la température de transition était immédiatement liée avec la quantité de lacunes dans l'état principal de la molécule [6]. Nos recherches montrent que le point critique dans la transition structurelle entre la structure idéale et la structure avec lacunes est aussi critique pour la transition thermodynamique [5]. Dans la région des structures idéales avec $p > p_c$, seul l'état en fusion est possible, quel que soit la température du polymère. Fig. 5 montre la diagramme de phase sur le plan (T, p).

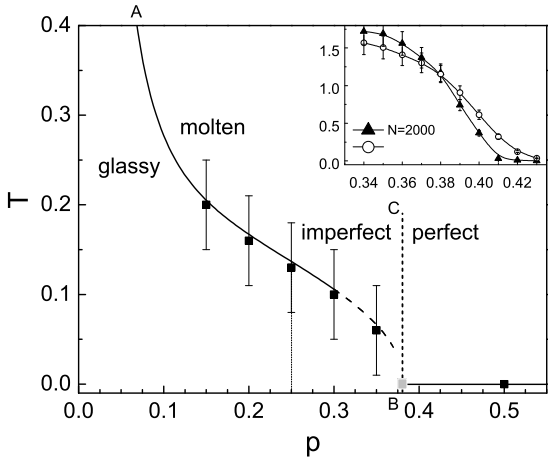


Figure 5. Diagramme de phase illustrant la corrélation entre la transition de phase thermodynamique et la transition structurelle.

3 Modèle d'intervalles aléatoires de l'ARN

Une nouvelle approche a été proposée pour la description de la structure secondaire du polymère pareil à l'ARN dans les termes d'un problème de transport d'optimisation. Un nouvel algorithme mathématique pour la détermination de cette structure a été obtenu dans le cadre du modèle d'intervalles aléatoires. Dans ce modèle, l'énergie d'interaction des monomères $\varepsilon_{i,j}$ est comptée comme la fonction convexe de distance entre les polymères le long de la chane. d'un point de vue physique, l'interaction électrostatique $\sim 1/d_{i,j}$ peut être un exemple d'une

telle interaction. Dans notre modèle, nous admettons

$$\varepsilon_{i,j} = u \ln |x_i - x_j|; \qquad (j \neq i)$$

où u est quelque valeur positive, et x_i, x_j sont les coordonnées des monomères i et j le long de la chane. Les distances $d_i = |x_{i+1} - x_i|$ entre les monomères voisins suivent quelque distribution $P(d_i = d)$.

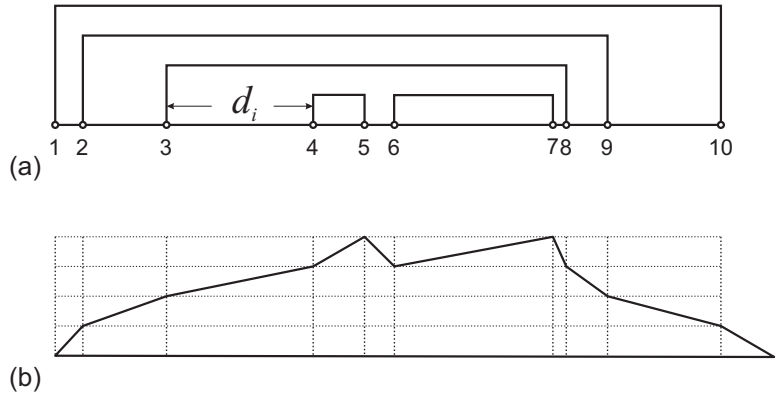


Figure 6. Modèle d'intervalles aléatoires du polymère pareil à l'ARN: présentation en arc (a), et chemin correspondant de Dyck (b).

[7] démontre que dans ce modèle l'énergie de l'état principal F répond à la relation récursive:

$$F_{i,i+k} = \min\left[\varepsilon_{i,i+k} + F_{i+1,k-1};\ F_{i,i+k-2} + F_{i+2,i+k} - F_{i+2,i+k-2}\right]$$

Ainsi, l'énergie $F_{i,i+k}$ pour le potentiel convexe d'interaction entre les monomères satisfait non seulement l'équation non-locale standard mais aussi l'équation locale possédant des propriétés de sous-additivité et de sous-modularité. l'ouvrage donne une analyse détaillée numérique et analytique de deux distributions d'intervalles $P(d_i = d)$: distribution de Gauss et distribution exponentielle. Il a été démontré quune transition topologique entre l'interaction consécutive et la configuration encastrée avait lieu dans le modèle d'intervalles aléatoires du polymère pareil à l'ARN pour la distribution Gauss. Le paramètre qui contrle le crossover est la valeur de dispersion dans la distribution de Gauss $f(d, \sigma)$. Dans la distribution exponentielle $f(d, \gamma)$ où la probabilité de longs intervalles nest pas exponentiellement petite, la présence de telles queues lourdes mène à une autre conduite structurelle en fonction du paramètre de distribution γ. La distribution exponen-

tielle est caractérisée par la présence du maximum de la fonction de la dimension de la structure $\langle h(\gamma)\rangle$ si $\gamma = 1$. l'expérience numérique est bien accordée avec les estimations analytiques [8].

4 Analyse statistique des réseaux

4.1 Analyse des amas

L'analyse des amas sur les réseaux est une tche importante. Premièrement, cette analyse d'amas réels (génétiques, d'expression) permet de faire des conclusions sur leur fonction biologique. Deuxièmement, cette analyse est utile dans le cadre des problèmes de prédiction d'amas sur le réseau arbitraire. Au cours du travail, nous avons étudié des amas obtenus lors de différentes expériences sur le réseau WormNet [9] y compris toutes les interactions connues entre les gènes *C.elegances*. Pour l'étude, on envisage utiliser l'analyse des plus courts chemins dans ces amas. La cohésion d'un amas, à l'égard des plus courts chemins, est déterminée comme suit

$$k_{cl}^{SPF} = \frac{2}{N(N-1)} \sum_{i,j=1}^{N} \frac{1}{d_{i,j}}$$

Pour la prédiction de régulateurs potentiels (), nous avons appliqué des coefficients à l'amas

$$k_{M}^{SPF} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{d_{i,M}}$$

Ici, N est la dimension de l'amas (nombre de gènes). Fig. 7 montre comment cette analyse peut être utilisée pour la validation de l'amas sur le réseau. La dépendance de cohésion spécifique est donnée pour les amas eQTL. A titre de comparaison, la dépendance de cohésion classique pour ces amas est présentée. Les lignes représentent les dépendances correspondantes pour des amas aléatoires sur le réseau. La grande différence entre les amas aléatoires et expérimentaux dans l'analyse SPF (à la différence de l'examen classique) permet de valider ces amas sur le réseau. l'analyse des régulateurs potentiels à l'amas observé dans les expériences de choc thermique a permis de dégager plusieurs gènes entranés aussi dans le cycle déterminant la longévité d'un organisme (Fig. 8)[9].

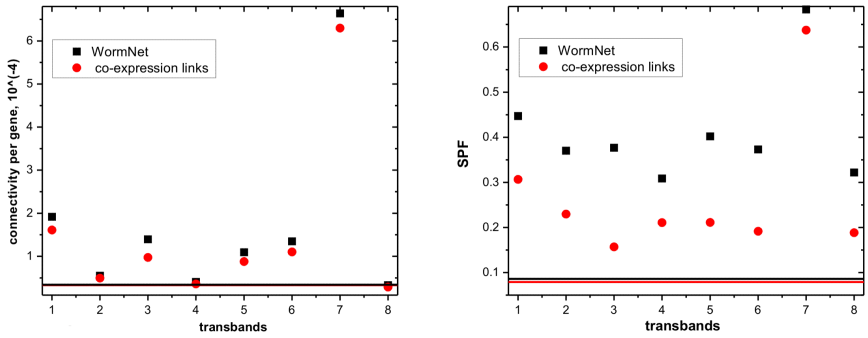


Figure 7. Dépendances de cohésion spécifique dans une description classique (à gauche) et en termes des plus courts chemins pour huit eQTL hotspots sur le réseau entier (noir) et sur la co-expression du sous-réseau (rouge). Les lignes correspondent aux dépendances pour les amas aléatoires dans le réseau.

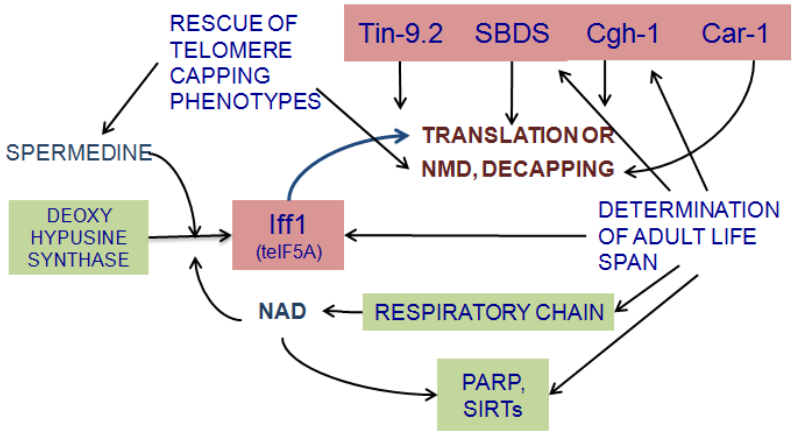


Figure 8. Schéma potentielle de la fonction des régulateurs prédits dans le cycle déterminant la longévité *C. elegance*.

4.2 Réseaux aléatoires

Il a été réalisé une analyse détaillée de la distribution des motifs pour les réseaux aléatoires d'ErdósRényi. Les motifs sont les petits sous-graphes du réseau qui déterminent la structure locale du graphe. Les auteurs [10] ont proposé d'analyser la fréquence d'apparition de motifs dans le réseau étudié par rapport à la fréquence randomisée. La procédure de randomisation du réseau réside en commutation de paires de ctes choisies au hasard (voir Fig. 9). Chaque commutation change les concentrations des sous-graphes à trois sommets. Les commutations sont décrites par des réactions élémentaires . On peut vérifier immédiatement quune seule

réaction élémentaire non-triviale existe dans un réseau non-orienté:

$$[0] + 3[2] \rightleftarrows [3] + 3[1]$$

Nous discutons la dynamique du réseau dans le champ extérieur h. Le potentiel

undirected subgraphs-triads				
symbol	[0]	[1]	[2]	[3]
concentration	c_0	c_1	c_2	c_3

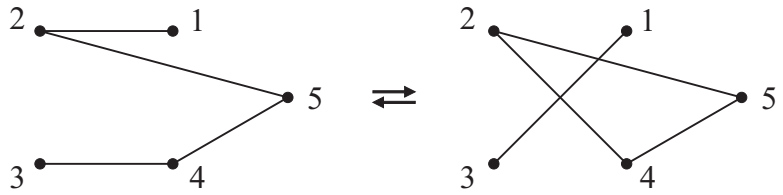
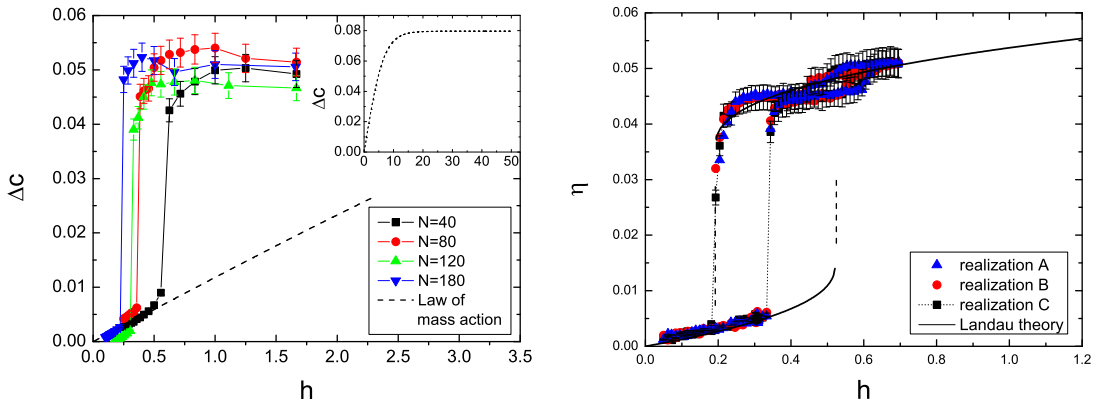


Figure 9. Motifs possibles dans un réseau non-orienté (figure en haut) et exemple de commutation singulière: au lieu de la paire de liaisons (1,2) et (3,4), les liaisons (1,3) et (2,4) apparaissent, avec cela, les valences des sommets sont préservées.

attirant vers l'un des motifs (motif 3) signifie que l'énergie du système à chaque pas de la randomisation change comme suit

$$\Delta E = -\frac{1}{2}h(\Delta C_3 - \Delta C_0)$$

où ΔC_0 et ΔC_3 sont les changements de la quantité de triades de type 0 et 3 après le pas de randomisation. Et l'évolution en l'état d'équilibre du système dans le champ extérieur h peut être décrite à l'aide de l'algorithme de métropolis, admettant la possibilité du pas menant à la réduction d'énergie comme égal à 1 et égal à $e^{-\Delta E}$ dans le cas inverse. Nous avons montré que, pour les petits champs h, la dynamique du réseau était bien décrite par la loi de masses actives découlant de l'équation de réaction (Fig. 10). Avec les grands champs h, un changement par à-coups de la concentration du motif de Fig. 10 a lieu. Ce saut est conditionné par une nette croissance de l'entropie avec augmentation de la concentration du motif et se rapporte aux transitions de phase du premier genre. La transition observée peut être décrite dans le cadre de la théorie phénoménologique des transitions de phases de Landau [11].



(a) (b)

Figure 10. Distribution de motifs $\Delta c(h) = c - c(h = 0)$ dans le réseau aléatoire avec $p = 0.35$ dans un champ extérieur différent h. Les expériences numériques sont réalisées pour les réseaux aléatoires de la dimension $N = 40, 80, 120, 180$ le trait pointillé est la courbe analytique prédite par la loi des masses actives. Le graphique additionnel montre la loi des masses actives dans la région des grands champs h. La comparaison des données expérimentales $\Delta c(h)$ avec solution analytique avec les paramètres $\chi = 85$; $b = 4.55 \times 10^3$ $g = 6.5 \times 10^4$.

5 Conclusions

1. Un algorithme basé sur la programmation dynamique est élaboré pour la détermination de l'énergie et de la structure des complexes de liaison de deux polymères linéaire et pareil à l'ARN.

2. Les propriétés statistiques des complexes aléatoires linéaires et pareils à l'ARN sont étudiées.

3. La conduite critique de la structure du polymère aléatoire de l'ARN est étudiée numériquement et analytiquement en fonction de l'alphabet utilisé. Il a été démontré quil existait deux régions: pour les alphabets $c < c_c$, une structure secondaire idéale sans délétions se forme en résultat du folding, alors que pour $c > c_c$ la structure secondaire contient la dose finale de monomères libres. l'estimation analytique du point de transition $c_c = 2.87$ est proche de celle observée dans l'expérience numérique $c_c = 2.67$.

4. Une nouvelle approche est élaborée, empruntée du problème de transport d'optimisation, de la détermination de la structure et de l'énergie du polymère pareil à l'ARN dans le cadre du modèle de potentiel convexe d'interaction entre les monomères de séquence primaire. La statistique de polymères

pour différentes distributions de distances entre les monomères à l'intérieur de la séquence secondaire est étudiée numériquement et analytiquement.

5. Des nouvelles approches de l'analyse des amas sur le réseau sont proposées. Sur l'exemple de WormNet est des amas obtenus expérimentalement, l'utilisation des méthodes élaborées pour l'analyse de données biologiques est démontrée.

6. Les distributions des motifs de réseaux aléatoires d'ErdsRényi sont étudiées. La présence de la transition de phase de premier genre dans l'espace des motifs est démontrée, et la description analytique de la transition dans le cadre de la théorie phénoménologique du champ moyen de Landau est donnée.

References

[1] S. K. Nechaev, M.V. Tamm and O.V. Valba, (2011) "Statistics of noncoding RNAs: alignment and secondary structure prediction", *Journal of Physics A: Mathematical and Theoretical*, Vol. 44. No.19.

[2] P.-G. de Gennes, (1968) *Statistics of branching and hairpin helices for the dAT copolymer Biopolymers*, Wiley Blackwell (John Wiley & Sons), 6, 715.

[3] O.V. Valba, M.V. Tamm and S. K. Nechaev, (2012) "New Alphabet-Dependent Morphological Transition in Random RNA Alignment", *Physical Review Letters*, V. 109(1):018102.

[4] S. K. Lando, (2007) *Lectures on generating functions*, MCCME.

[5] A.Y. Lokhov, S.K. Nechaev, M.V. Tamm, O.V. Valba, (2013) "New phase transition in random planar diagrams and RNA-type matching", arXiv:1307.2170.

[6] R. Bundschuh, T. Hwa, (2002) "Statistical mechanics of secondary structures formed by random RNA sequences" *Physical Review E*, V. 65.

[7] J. Delon, J. Salomon, A. Sobolevski, "Local Matching Indicators for Transport Problems with Concave Costs", (2012) *SIAM Journal on Discrete Mathematics*, 26, 801-827.

[8] S.K. Nechaev, A.N. Sobolevskii, O.V. Valba, (2013) "Planar diagrams from optimization for concave potentials", *Physical Review E*,Vol. 87, No. 1.

[9] O.V. Valba, S.K. Nechaev, M. G. Sterken, L.B. Snoek, J. E. Kammenga, O. Vasieva, (2013) "On predicting regulatory genes by analysis of functional networks in C.elegans", arXiv:1302.3349.

[10] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, (2002) "Network motifs: simple building blocks of complex networks" *Science*, 298, 824-827.

[11] V.A. Avetisov, S.K. Nechaev, A.B. Shkarin, M.V. Tamm, O.V. Valba, (2013) " Islands of stability in motif distributions of random networks", arXiv:1307.0113.