INSTITUT DE FRANCE
Académie des sciences

Réception des Associés étrangers élus en 2005 / 12 décembre 2006

INFORMATIQUE ET BIOLOGIE MOLÉCULAIRE : NAISSANCE D'UNE NOUVELLE DISCIPLINE

Michael WATERMAN
Professeur de sciences biologiques, de mathématique et d'informatique
Université de Southern California (États-Unis)

Au début des années 70, la biologie moléculaire avait commencé sa longue ascension vers la prééminence. La double hélice était connue depuis 20 ans et les séquences d'un certain nombre de protéines et d'acides nucléiques avaient déjà été déterminées. Le premier article sur la séquence d'un ARN de transfert avait présenté la célèbre structure en feuille de trèfle, une réussite brillante d'analyse de séquence (1). Margaret Dayhoff avait publié plusieurs volumes de son *Atlas of Protein Sequence and Stucture*. Ce fut le point de départ pour plusieurs d'entre nous, qui commençaient à travailler dans ce domaine, lequel sera connu, 25 années plus tard, sous le nom d'« application à la biologie des méthodes mathématiques d'analyse » ou « Bioinformatique » (2). Ces volumes contiennent les séquences protéiques et nucléiques connues à l'époque. M. Dayhoff y a également inclus des articles provocateurs sur des alignements de séquences marquants et la notion de familles de gènes. Cependant, l'évènement qui a fait exploser le domaine d'analyses mathématiques des séquences fut l'invention des techniques rapides de séquençage de l'ADN par Sanger à Cambridge et par Maxam-Gilbert à Harvard. Les méthodes informatiques deviennent alors critiques pour lire les séquences elles-mêmes et l'habituelle analyse « à l'oeil » de ces séquences, dont le volume s'accroissait sans cesse, devenait, à la fois, de plus en plus imprécise et de plus en plus irréaliste.

La comparaison de séquences a été ma porte d'entrée dans ce domaine. En 1974, je rejoins Temple Smith et Bill Beyer à Los Alamos pour un été passionnant. Je ne connaissais pratiquement rien en Biologie et, comme j'étais attiré par le sujet, je voulais lire, dans les bistros de Santa Fe, le petit livre de Davidson, *The Biochemistry of Nucleic Acids* (3). Les auteurs de la nouvelle édition nommaient ce livre «Guide des acides nucléiques pour enfants » et je suis devenu accro. Les êtres humains apprennent à lire des textes linéaires et sont attirés par leur analyse et leur modélisation. Que notre texte génétique puisse être lu de manière linéaire me paraissait stupéfiant. Heureusement, David Sankoff et Peter Sellers avaient déjà commencé un travail rigoureux sur l'alignement de séquences (4). Cet été là, nous avons élargi leur travail pour inclure des espaces (insertions et délétions) d'une longueur supérieure à un nucléotide, ainsi que la comparaison simultanée de nombreuses séquences. Notre méthode pour résoudre ce dernier problème était très coûteuse en temps de calcul mais nous pensions déjà aux problèmes qui restent importants actuellement (5). Nous avons aussi

écrit un article sur l'évaluation des arbres évolutifs, en partant des données de séquences du cytochrome C. Si mes souvenirs sont exacts, nous en possédions 27, ce qui nous paraissait une somme de données presque écrasante. C'était vrai à l'époque (6).

Quand les introns ont été découverts, il a fallu trouver une autre approche pour les alignements de séquences. Smith et moi-même avons presque immédiatement commencé à créer une méthode permettant de faire face à ce nouveau problème que nous ne savions pas formuler de manière précise. La solution, avec un énoncé clair de la nature du problème, est arrivée deux ans plus tard sous la forme de ce qui est devenu l'algorithme de Smith et Waterman (7). Cette méthode est au coeur de presque toutes les méthodes de recherches des bases de données, y compris le BLAST, qui peut être considéré comme une heuristique très rapide de notre algorithme. Il y a une autre question-clé liée aux comparaisons de séquences en général : la signification statistique. Avec quelle probabilité peut-on trouver un certain pattern à partir de séquences aléatoires, sans relations biologiques ? On peut faire de telles estimations à chaque tour de BLAST et notre travail sur l'approximation de Poisson propose une partie de la solution, qui fournit ces estimées.

Au début des années 80, j'ai décidé d'étudier le problème fascinant des patterns de régulation des régions promotrices. Ces patterns sont souvent courts et n'apparaissent presque jamais de manière précise. De plus, dans une quelconque comparaison de séquences promotrices, on ne trouve presque jamais le véritable pattern. J'ai utilisé une approche qui notait chaque pattern possible par l'approximation de son occurrence dans chacune de nombreuses séquences (8). Plus tard, les méthodes d'apprentissage statistique sont devenues courantes et ainsi, après 15 années, ma méthode, actuellement appelée méthode combinatoire, fut de nouveau utilisée.

Le repliement des ARN est une énigme intéressante et j'ai été l'un des pionniers en ce domaine, en utilisant les aspects énergétiques afin de calculer le minimum d'énergie libre pour de telles structures, en perturbant les comparaisons de séquences (9). Des implémentations variées sont utilisées aujourd'hui. La solution implique l'énumération des configurations potentielles et ce type de mathématiques a été employé depuis par de nombreux chercheurs. J'ai aussi utilisé une approche « combinatoire » pour trouver le repliement consensus d'une famille de molécules d'ARN.

Dans le projet Génome Humain et dans les programmes de grand séquençage, le génome est couvert par des fragments chevauchants aléatoires ou clones. Eric Lander et moi-même avons étudié ce processus stochastique et calculé les estimées de couverture en fonction du nombre et de la longueur des clones, avec la technique de chevauchement. La formule de Lander-Waterman s'est révélée fort utile pour ces projets et d'autres (10).

Plus récemment, mon équipe et moi-même nous avons travaillé sur les variations génomiques dans les populations humaines et autres (11, 12). Les données sur le polymorphisme nucléotidique ont apporté de nouvelles percées sur la structure de telles variations dans le génome humain et peuvent, dans certains cas, fournir des marqueurs pour l'identification de gènes responsables de maladies.

Un des leitmotivs de mon travail (et celui d'autres chercheurs) en bioinformatique est que technologie et biologie montrent la voie. Il n'y a pas de grande théorie en génomique, bien que chacun espère des découvertes qui seront universelles. Cependant, chaque nouvelle

technique semble créer de nouveaux défis informatiques et statistiques en sorte que les perspectives biologiques puissent être déverrouillées par la technologie

---

INFORMATICS AND MOLECULAR BIOLOGY : BIRTH OF A NEW DISCIPLINE

By the early 1970s, molecular biology had begun its long ascent to prominence. The double helix had been known for 20 years and there were a number of protein and nucleic acid sequences already determined. The first paper with a RNA sequence contained the celebrated cloverleaf structure, a brilliant feat of sequence analysis. (1) Margaret Dayhoff had published several volumes of her Atlas of Protein Sequence and Structure which were the starting place for several of us who began to work in the subject which was 25 years later to be known as computational biology or bioinformatics. (2) These volumes contained the protein and nucleic acid sequences which were known at that time. She also included provocative articles about scoring alignments and sequence or gene families.

But the events that made the area of computational sequence analysis explode were the invention of Sanger in Cambridge and of Maxim-Gilbert in Harvard of rapid DNA sequencing techniques. Computational methods were critical to read the very sequences themselves and routine "eyeball" analysis of the increasing volume of nucleic acid sequences was increasingly inaccurate as well as impractical.

Sequence comparison was my entry point into this subject. In 1974, I joined Temple Smith and Bill Beyer at Los Alamos for an exciting summer. I knew almost no biology and as I became pulled into the subject I would read in Santa Fe coffee shops the small volume by Davidson The Biochemistry of Nucleic Acids. (3) The authors of the revision refer to it as "The Child's Guide to the Nucleic Acids" and I was hooked. Humans learn to read linear texts and are drawn to model and analyze texts. That our genetic script could be read as linear text seemed quite amazing to me. Fortunately David Sankoff and Peter Sellers had already begun rigorous work on sequence alignment. (4) That summer we extended their work to include gaps (insertions or deletions) of length greater than one as well as comparison of many sequences at once. Our method for this last problem was quite costly in computer running time but we were thinking about problems which remain important today. (5)

We also wrote a paper on estimating evolutionary trees from cytochrome C sequence data. We had as I recall 27 such sequences which seemed an almost overwhelming amount of data, which was true at that time. (6) When introns were discovered it required a different view of alignment. Smith and I almost immediately began trying to create an alignment method to cope with this new problem for which we had no precise formulation. The solution, along with a clear statement of what the problem was, came two years later with what has become known as the Smith-Waterman algorithm. (7) This method is at the heart of almost all database search methods including BLAST which can be considered a very fast heuristic for our algorithm. There is another key issue related to sequence comparisons in general, which is statistical significance. How likely is a found pattern to appear from random,

biologically unrelated sequences? Such estimates are made with every run of BLAST and our work on Poisson approximation is part of the solution which provides the estimates.

In the early 1980s I decided to study the puzzling problem of finding regulatory patterns in promoter regions. The patterns are often short and almost never appear exactly. Moreover in any comparison of two promoter sequences the true patterns can almost never be found. I used an approach which scored each possible pattern by its approximate occurrence in each of many sequences. (8) Later statistical learning methods came into prominence and then, over 15 years later, my method which now called a combinatorial method was again being used. RNA folding is an intriguing puzzle and I pioneered using energy functions to compute the minimum free energy structures using a perturbation of sequence comparison. (9) Various implementations are used today. The solution involves enumerating the potential configurations and that mathematics was taken up my numerous people. Also I took another "combinatorial" approach to find consensus folding of a family of RNA molecules. In the Human Genome Project and large sequencing projects, the genome is covered by overlapping random fragments or clones. Eric Lander and I studied this stochastic process and calculated estimates of coverage as a function of the number and length of the clones, along with the clone overlap technique. The Lander-Waterman formula proved useful in these and other projects. (10)

More recently I and my group have been working on data for population variation in the human and other genomes. (11, 12) The single nucleotide polymorphism data has provided new insights into the structure of such variation in the human genome and in some cases can provide markers for disease gene identification. One of the themes of my work and others in bioinformatics is that technology and biology lead the way. There is no grand theory for genomics, although everyone hopes for discoveries which are universal. However each new technology seems to create new computational and statistical challenges in order to unlock to biological insights to be provided by the technology.

References

1. Holley, R.W. et al, Structure of a ribonucleic acid, *Science* 147, 1462, 1965
2. Dayhoff, M.O. *Atlas of Protein Sequence and Structure*, 1965, 1966, 1967, 1969, 1972
3. Davidson, J.N. *The Biochemistry of Nucleic Acids*, Academic Press, 1976
4. Sellers, P. *SIAM J. Appl. Math.*, 26, 787, 1974.
5. Waterman, M.S., Smith, T.F. and Beyer, W.A., Some biological sequence metrics, *Adv. Math.*, 20 367, 1976
6. Waterman, M.S., Smith, T.F., Singh, M. and Beyer, W.A., Additive evolutionary trees, *J. Theor. Biol.*, 64 199, 1977
7. Smith, T.F. and Waterman, M.S., Identification of common molecular subsequences, *J. Mol. Biol.*, 147 195, 1981
8. Galas, D.J., Eggert, M. and Waterman, M.S., Rigorous pattern recognition methods for DNA sequences: analysis of promoter sequences from *E. coli., J. Mol. Biol.*, 186 117, 1985
9. Waterman, M.S., Secondary structure of single stranded nucleic acids, *Adv. Math. Suppl. Stud.*, I 167, 1978
10. Lander, E.S. and Waterman, M.S., Genomic mapping by fingerprinting random clones: a mathematical analysis, *Genomics*, 2 231, 1988
11. Zhang,K., Deng, M., Chen, T., Waterman, M.S., and Sun, F., A dynamic programming algorithm for haplotype block partitioning, *Proc. Natl. Acad. Sci. USA*, 99 7335, 2002
12. Valouev A, Schwartz DC, Zhou S, and Waterman MS., An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proc Natl Acad Sci U S A* 2006