Updated November 23rd, 2015

CONTRIBUTIONS TO BIOINFORMATICS

JACQUES NINIO

(included in the web site http://www.lps.ens.fr/~ninio)

TOPICS DISCUSSED HERE:

OVERVIEW

INTRODUCTION

I started programming in 1967 for practical purposes. One was the exploitation of protein sequence data that were becoming available one by one, in relation to the genetic code, the other was for solving practical problems in relation to the exploitation of X-ray data, more precisely, the data that I was collecting on « small angle X-ray scattering of transfer RNA in solution » – my thesis work. In this domain, I had rather astute algorithmic ideas, including one that led me later, to the algorithm for fast comparison of nucleic acid sequences [1], that is the motor of the famous « BLAST » engine [2]. I also designed an (obvious) deconvolution algorithm that was later re-invented and became the heart of the famous « retro-propagation » algorithm in cognitive sciences [3]. At that time, computer work was considered as « cooking ». We published the results, not the computer methods. Later, when I moved to Institut Jacques Monod, I teamed with Jean-Pierre Dumas on the project of predicting secondary structures in RNA molecules. We did fine algorithmic work. However, my main motivation was to derive a set of free energies that might apply to non-standard base pairs – not only the so-called wobble G.U pair but also to all kind of « odd » interactions : G.A, U.U, etc. At that time, people did not believe that these interactions could occur within nucleic acid stems, yet we could assign energy values not too far from the energy values of the G.U pairs. The results were taken seriously, but the method used a « tolerable noise principle » (see below) that is not yet assimilated by the bioinformatics community. In parallel with the simulation and algorithmic work on RNA structures, I developed computer graphic skills in relation to stereoscopic images (see the web chapter on stereo vision).

In 1981, I organized the first international bioinformatics meeting (see below, and announcement poster in the end), then kept working a little on secondary structure algorithms, and on graphical representations of sequences. In 1989 I published with Eduardo Mizraji an epistemological work that stands as my swan song in the field of bioinformatics. One of my motivations in bioinformatics was to interact with experimentalists who were producing heaps of sequences around me, and help them answering the questions arising from the sequences. Unfortunately, the questions turned out to be conceptually very poor. I became disenchanted. I turned the page, and became involved in visual perception.

COMPUTER WORK AT GIF-SUR-YVETTE

I started my thesis work in 1964, in Vittorio Luzzati's laboratory (Centre de Génétique Moléculaire, CNRS, Gif-sur-Yvette, France) on the structure of transfer RNA, as probed by the small-angle X-ray scattering technique (see section on RNA structure). Computers were beginning to become auxiliary laboratory tools. We had an access to a computer centre at Orsay, running a UNIVAC machine. Programs in FORTRAN were carefully written and typed, line by line on punched cards. There was, in Luzzati's laboratory a professional programmer, Cora Saludjian, wife of Pedro Saludjian, a biophysicist with a broad culture. Both were immigrants from Argentina. Cora's life ended tragically in 1967, after giving birth to a son, Luca. After her death I started programming myself.

The correspondence between codons and amino acids in the genetic code was completely elucidated in 1965, and this opened the way for phylogenetic studies in evolution. In parallel with my structural work on transfer RNA, I started playing with the haemoglobin and cytochrome c protein sequences that were then becoming available. I devised a measure of the distance between two sequences that was not a sum of the amino acid changes, but a ratio of the changes that could not be explained by single nucleotide substitutions to those who could. All the programming was done by Cora Saludjian. The ratios defined above appeared to cluster around a few discrete values. This work never came close to being published (see forthcoming section on molecular evolution).

In interpreting the X-ray data, we were confronted with two computational problems. I solved them by devising algorithms that have interesting stories (see "retro-propagation algorithm" and "a 3d sorting algorithm" below).

COMPUTER WORK AT INSTITUT JACQUES MONOD IN PARIS

In 1976, at Institut Jacques Monod,I started to have frequent contacts with a laboratory technician trained in chemistry and biochemistry, Jean-Pierre Dumas. Dumas had a wide curiosity, and an eagerness to learn far beyond what his official degrees could suggest. He first came to me to discuss points of Arabic language, that he was studying at that time, and in which I was competent, having spent all my childhood in Egypt. He then started to acquire competence in computer programming, to the point of being able to write programs in machine language, and even write compilers. He also acquired excellent knowledge of the inner workings of the computers. One day he came to me asking if I could suggest a biologically relevant problem for which he could write a computer program. I immediately proposed the prediction of RNA secondary structure. (See "Contributions to RNA secondary structures" below).

The interest for computer methods to deal with nucleic acid sequences was groing rapidly. There was in my laboratory a young student Philippe Marlière, who came with many ideas on problems to be addressed (in particular he was interested in algorithms for optimal simultaneous comparisons of sequences) and he had good contacts with the people in charge of bioinformatics at the Pasteur Institute. Close to us, in the Jussieu campus, there was a whole team, headed by Jean-Sallantin, coming from the field of artificial intelligence. This team tried to apply the "expert system" approach to biological problems. We had frequent contacts with them. For many years, I had, in the topic of bioinformatics frequent, regular contacts with the French people working in the field, and in particular, with the group headed by Richard Grantham in Lyon, who was working on molecular evolution. I was invited to organizational meetings, in particular those of the public bioinformatics platform (CITI2, headed by Philippe Dessen). I was generously funded by the French Agency for Informatics. I used most of the funds to buy graphic plotters that were subsequently used for my work on stereoscopic vision. With the benefit of the ascending currents, I organized the first international meeting on bioinformatics (see "EMBO meeting on pattern analysis in nucleic acid and protein sequences" below), and developed the algorithm for fast sequence comparisons at the heart of the «BLAST » package, and used everywhere to-day (see "fast sequence comparison algorithm" below)

In the course of my work on stereoscopic vision, I had acquired competence in graphic programming, and it is thus, very naturally, that I became involved in work on graphical representations of nucleic acid sequences (see "graphical coding of sequences" below). After that, I tried to take altitude, and did an epistemological work, with Eduardo Mizraji in the line of Minsky and Papert's Perceptron book (see "epistemological work : the limitations of pattern analysis" below). This was in 1989. After that, I did a lot of computer programming, mostly, interactive computer graphics, but in connection with my interests in visual perception. I thought that I could extend my competence in the stereo matching problem in vision to a molecular problem : the matching of bi-dimensional gels of proteins. I thought that I had the natural, obvious, solution to the problem, but did not find any one willing to send me experimental data to test my ideas.

RETRO-PROPAGATION ALGORITHM

When I was working on small-angle X-ray scattering in the 1960's, there was a conceptually simple problem, known under the mathematical term of "deconvolution". In the acquisition of X-ray data, one could use an apparatus producing a narrow X-ray beam (ideally, with a cross-section which could be assimilated to a point). Then, the theory relating the structure of the studied molecule to its X-ray diffusion pattern was straightforward. Or one could use X-rays coming through a wide slit. The beam could then be far more intense, but the signal would be blurred due to the dispersion of the rays hitting the sample. With an accurate physical description of the X-ray beams coming out of the slits, one can, in principle, recover the signal that would have been theoretically produced by a point-collimated source.

At that time, James Albert Lake (who became later well-known for work on ribosome structure and on phylogenetic trees) was working on the same subject as I (transfer RNA structure by small-angle X-ray scattering) and was in fact ahead of me, having published his results in 1968 [4]. In this work, Lake used a crude algorithm to deal with the deconvolution problem [5]. Let theor(x) be the (unknown) theoretical signal produced by a point-collimated X-ray beam, let obs(x) be the experimentally observed signal, and let shape(x) be the function describing the spread of the X-ray beams. Let us start with what we believe to be a first approximation to the theoretical function, say guesss1(x). Applying shape(x) over guess1(x), one obtained a function simul1(x) that differed substantially from obs(x).

Lake then constructed a second guess to theor(x) which was deduced from the first guess by simply multiplying guess1(x) at each abscissa x, by the ratio obs(x)/simul1(x). Unfortunately, this procedure does not guarantee the convergence of the successive guesses towards theor(x).

So, I had the (obvious) idea to correct guess1(x), at EVERY point that contributed to a discrepancy between simul(x) and obs(x). I wrote a program doing that, described it in my thesis ([6], pages 23-24) and it was used subsequently by other workers in Luzzati's laboratory.

In essence, this algorithmic procedure, in which (i) one has successive guesses, producing successive simulated outputs, (ii) one compares the simulated output to a target output and (iii) one corrects the guess at all points which contributed to the discrepancy between simulated and target outputs, is now famous under the name of "learning by back-propapagation" or learning by "retro-propagation". Rumelhart, Hinton and Williams, 1986 [3] acquired great fame in cognitive sciences for their proposal that neural networks make use of such a procedure.

A 3-d SORTING ALGORITHM

At the time of my thesis, the problem of predicting the theoretical X-ray diffusion curve from a molecular model was computationally heavy, even in the very simple case with which I was concerned, small-angle X-ray scattering. If a molecule is composed of N atoms, then the theoretical X-ray scattering curve can be deduced simply from a function of the positions of the atoms, called the Patterson function. Things are actually more difficult because what matters is the difference between the molecule's local electron density and the electron density of water. Vittorio Luzzati and I found a solution to the problem, still used or discussed to-day (see, e.g.,[7, 8]). I had the computational idea, and Luzzati made it physically correct. The solution was to divide the space into cubes, having about the same volume as that of a molecule of water, determine the electronic content of each cube (adding the number of electrons of all the atoms inside the cube), subtract the water electrons, put the results at the centre of the cube, compute the Patterson function for the fictitious molecule made of electrons at the centres of the cubes, and correct for the sampling effect (this last point was Luzzati's essential contribution). The modern addition to this method was to make more precise the contribution of the shell of water molecules surrounding the macromolecule, for instance by rolling a sphere of « dense water » over the surface of the macromolecule [7].

In practice, taking the x, y, z coordinates of an atom in the molecular model, I constructed an index which was a single integer, such that all atoms inside a cube had the same index, from which the coordinates of the centre of the cube could be recovered. Thus, the 3d coordinates were converted into a set of integers, each one affected with a particular electron density. The Patterson function was then easily computed. This algorithm is mentioned elliptically in my thesis [6], bottom of page 28. In the JMB paper [9], the physical foundations are discussed, but there is not a single word about the algorithm. It must be said that, in these times, much of the algorithmic work was considered as "cooking", and noble science was in the physics.

When, about 10-15 years later I became involved in the bioinformatics of sequence analyses, I made the link with the problem of partitioning the atoms into cubes that were indexed by single integers (see the section below on fast sequence comparison algorithms).

PREDICTION OF RNA SECONDARY STRUCTURES

While the prediction of protein structures had been for two decades under massive attack, the prediction of RNA secondary structure was still in 1976, for reasons unknown to me, in infancy. There was a ridiculous paper by Pipas and McMahon in PNAS [10], showing extremely poor results on such a simple molecule as transfer RNA. My main scientific interest in the work was not algorithmic. I was more interested in the energetics of nucleotide pairings, and especially the energetics of non-Watson-Crick pairs. I thought that if we had a powerful tool for generating RNA secondary structure patterns, we could find the set of energy values that led to the best predictions, and from there, perhaps, derive novel insights into biologically important problems, such as codon-anticodon recognition. Prior to this work, people were using sets of energy values obtained from sparse biophysical experiments. These values could not even lead to the cloverleaf folding in transfer RNA.

So, Jean-Pierre Dumas and I took the problem from the two sides: improving the algorithms for generating RNA secondary structures, and refining the set of energy values that were fed into the calculations. In the end we were able to propose two complete sets of energy values that allowed a prediction of the cloverleaf structure in 90% of the cases. [11]. Jean-Pierre Dumas, who, at that time, was in rebellion with the institution, decided very firmly not to sign the paper.

There were, a few years later, claims by other authors that they had achieved better success, but they were in fact using a less stringent criterion of success. While we were giving the success rates for obtaining correctly the complete structures, the success rates used by others were success rates for obtaining the individual stems of the cloverleaf structures. So, their 90% success probability corresponded roughly to our 90 per cent to the power four, or 66% probability !

The 1979 Biochimie paper [11] was well quoted, as an empirical work proposing a set of energy values, in particular for features for which detailed data were needed (e.g., the penalties for bulge loops as a function of their lengths), but its real conceptual importance was missed. There were in fact two breakthroughs in this paper.

First, the very fact of considering all non- Watson Crick pairs as legal pairs, to which energy values should be assigned was rather bold at that time. If one reads again the articles published at that time, one will see that people were thinking of secondary structures involving only G.C and A.U pairs, or, for the most advanced of them, structures also including the wobble G.U pairs. Odd pairs like G.A or U.U were simply considered to be looping out. Now, odd pairs are receiving serious consideration (see, e.g., the review by Leontis and Westhof [12], but they are also considered sometimes as « mismatches » without precision about their structures [13]).

Second, and this is highly relevant to bioinformatics, the method that was used to derive the energy values for odd pairs (and accessorily, for other non standard features) was an entirely original method, with deep implications. The method is about hidden constraints and tolerable noise. Up until now, even 35 years later, the scientists in the field do not seem to have noticed the method, or seized its potential importance. It is explained further down in the section «Looking for hidden information in nucleic acid sequences (the tolerable noise principle)».

Using the tolerable noise procedure, we were able to derive a set of energy values for all the odd pairs, although they are rare in the tRNA stems [11, 14]. The reward came when we applied the energy models to structures in which odd pairs take part, and so cannot be underestimated. Odd pairs are frequent in 5S RNA structures, and when we extended our secondary structure explorations to the 5S RNA sequences, there was very little to change in the set of odd pairs stabilities [14]. This was extremely gratifying. The set of energy values for the odd pairs was later refined further [15], taking into account the observations made by Woese et al. [16] and Traub and Sussman [17] on the occurrence of the G.A odd pair at the entrance of internal loops. The article in Biochimie 1985 [15] marked the end of my work on secondary structures. I did not have the collaborators to go beyond this point. It was about time for Abou-ela, Koh, Tinoco and Martin to publish their own set of energy values for the odd pairs [18] derived, in principle, from their own biophysical measurements. However, their paper seemed to me to be substandard, and I wonder whether or not the authors would have been so confident about their experimental determinations, had they not had our own results in their hands. If one looks closely at another article by Ignacio Tinoco and co-workers [19], one has difficulties determining the origin of some data given there as obtained from biophysical measurements. They mention using « free energies that are different from those used earlier ». It seems to me that some of the experimentally missing values had been chosen by analogy with our own data sets.

THE INCOMPATIBILITY ISLETS ALGORITHM

The algorithms we developed were briefly presented in [1] and, in more detail, but in French, in [15]. Ideas for further developments of the field were also presented in [15]. The generation of RNA secondary structures proceeded in three steps. First, we built an exhaustive list of all the potential secondary structure stems that could take part in a structure, then we constructed for each segment, a list of all the segments with which it could not, according to some criteria, be present simultaneously in a structure. Then we made a "branch and bound" tree search to construct all potential structures that satisfied the compatibility criteria, and retained the 5 or 10 best candidates. The search was speeded up using several tricks. One was in the exploitation of the incompatibility relationships. The potential segments were grouped into clusters named "islets" such that each segment in the cluster was incompatible with all other segments of the cluster. Then a valid solution contained at most one segment from each cluster. This was my idea, and I called the algorithm, the "incompatibility islet algorithm". Prior to our work, the fastest algorithm, according to Haralick and Elliott [20] was a bit parallel look ahead depth first forward checking algorithm. Our « incompatibility islets » algorithms was running much faster than the classical forward checking algorithm, although we cannot be confident that it would be also superior in other problems (e.g., solving a sudoku grid). In any event, a recent survey of the literature indicates that the incompatibility islets algorithm has not yet been rediscovered, and that current strategies, based for instance on the « dancing links » method advertised by Knuth look rather clumsy with respect to the incompatibility islets algorithm. But again, our algorithm may be optimally adapted to the molecular folding problem, and not applicable with benefit to other situations.

Jean-Pierre Dumas found the way to implement the tree search in an extremely efficient way. The compatibility relationships between a potential segment and all other segments were encoded as single bits, juxtaposed in 60 bits words. The construction of the RNA structures advanced by making logical AND operations on the 60 bits words. This is the « bit parallel » trick (See his PhD thesis [21]). Dumas also succeeded to fool the CDC supercomputer, making it believe that it was using much less memory than it really did. This allowed us to carry out a huge amount of work on a very limited budget. After Dumas' departure for the Salk Institute, where he worked on the comparisons of 2d protein gels, Manolo Gouy - a mathematician having switched to molecular evolution - joined my group for a postdoctoral period of one year. Manolo was very active in a bioinformatics project in Lyon, headed by Richard Grantham. Grantham's group was working on codon usage and on automatic clustering of protein and nucleic acid sequences. They had launched an ambitious data bank project (called ACNUC) and worked for years on the collection and the annotation of biological sequences. During the year he spent at the Jacques Monod institute, Manolo Gouy worked mainly on the incompatibility islets algorithm. He developed strategies for optimizing the partitioning of the segments into islets, and on the optimal order for including segments into a structure. He made the programs more user-friendly, and provided the canonical descriptions of the algorithms in [15].

Our programs, contrary to their contemporary Nussinov-Jacobson [22] and Zuker-Stiegler [23] competitors (that could run much faster and thus dealt with much larger sequences) could deal with extremely complex energy models, allowing every possible base-pair, and non additivities in the energies of the various components of a structure. The programs also provided, from the very beginning, not only an optimal structure, but all 5 or 10 (or any number fixed in advance) of sub-optimal structures. Statistics on the free-energy difference between the optimal structure and the next one were explicitly provided in [11]. Therefore, when Williams and Tinoco published their paper in [24] claiming right at the beginning of their abstract that previous programs "were able to predict only one optimal structure", this was a shameless lie that, unfortunately, induced other people in error, distracting their attention from our own work.

Folding of random sequences. The secondary structure program was used by Philippe Marlière, a young, brilliant and versatile evolutionist to explore the folding patterns in random sequences with various constraints [25]. I hoped to use the programs to explore the phylogeny of transfer RNAs. Having a couple of closely related sequences of tRNAs, can we reconstruct the evolutionary pathway linking one sequence to the other? Under the assumption that each of the intermediates along the pathway is functional, one can hope to eliminate many possible intermediates, and end up with an explicit evolutionary pathway, in which the intermediates may differ from those usually postulated (on the basis of consensus sequence arguments, or of minimal changes arguments). So, I took two old friends, the valine-specific tRNAs I and II from E. coli, that differ at 4 positions, and looked at all possible intermediate sequences. Unfortunately, too many intermediates had good cloverleaf structures, so the criterion was not discriminative enough and I abandoned this line of work. Other workers have since addressed the problem of the true evolutionary intermediates, with more determination than I (e.g., Lee, DSouza and Fox with wet molecular genetics [26], or Fontana and Schuster [27] with computer tools). To-day, I would challenge the basic assumption of this work - that most changes retained by evolution occurred as single step mutations. Indeed, I believe now that multiple changes are far more frequent than people think, and have shown how simultaneous multiple changes can occur in bacteria, through "transient mutators" [28] and how correlated mutations can occur in higher organisms through gene conversions [29].

FAST SEQUENCE COMPARISON ALGORITHM

As a prelude to the article describing secondary structure algorithms, I included a short section (57 lines) describing a fast algorithm for looking at sequence repetitions or homologies [1].

First step : The sequences are recoded in terms of overlapping oligonucleotide subsequences (« n-tuples »). For instance, there is an integer between 1 and 4096 assigned to each of the possible 4096 hexanucleotides. The nucleotide sequence is recoded as a sequence of overlapping hexamers. Ideally, we would like to know the positions of all the hexamers of each kind. A clumsy programmer would construct 4096 tables or so and fill each of the 4096 tables, with the positions of occurrences of the relevant nucleotide. However, there is a much more astute, economical way, of encoding the information, using merely two auxiliary one-dimensional tables !

Second step : We construct two auxiliary tables, T and M. Table T is exactly 4096 positions long, and Table M has the same length as that of the original sequence. We scan the recoded sequence just once and encounter for instance hexamer 1365 at position 827 of the sequence. We look at position 1365 in Table T and find, for instance, 775. This means that 775 was the last position at which hexamer 1365 was encountered. We then replace 775 in Table by 827 the new last position of the hexamer, and put 827 in position 775 of Table M. In this way any filled position in Table M (e.g., 775) indicates the position of next occurrence of the same hexamer.

Third step. With minor variations, this basic organization of sequence information can be adapted for various specific tasks, such as the search for repetitions, symmetries or homology between sequences.

At the beginning, you had sequences, and the indexes were a function of the position. At the end, you had a table T in which the positions are functions of the indexes. This type of switching between function and variable is well known in mathematics. Homologies are detected at once, looking at the indexes that corresponded to more than one sequence position. The algorithm was very easily developed, by analogy with my previous work on 3d sorting. The trick is so obvious that I never believed that it could not have been used before, in a different context. In any event, it was new in the field of biological sequences comparisons. It was rediscovered by several authors (e.g. Wilbur and Lipman [30], Karlin et al. [31] . It seems that the sequence homology searches BLAST and FASTA (e.g., Altschul et al. [2], used by tens of thousands of people around the world run along the same principles). I was well aware of the fact that looking for strict homologies was just an initial step, to be followed by a search for further, less strict homologies, beyond the points of strict homologies. (page 199 in [1], last three lines of the first paragraph). The people who « rediscovered » the Dumas-Ninio 1982 fast comparison algorithm were reluctant to admit our priority. Thus, Wilbur and Lipman wrote in the abstract of their paper [30] :

> « We present an algorithm for the global comparison of sequences based on matching k-tuples of sequence elements for a fixed k. The method results in substantial reduction in the time required to search a data bank when compared with prior techniques of similarity analysis, with minimal loss in sensitivity. ». This is an obvious lie, unless they mean something specific about data banks, not on the fast comparison algorithm. Otherwise, they write in their paper « to locate all k-tuple matches, we follow a method described by Dumas and Ninio [12] ». But they say nothing about the one-dimensional representation of the data described here in step 2.
>
> Karlin et al. [32] wrote :
> « a new high speed computer algorithm is outlined... » (summary).
> « Currently available programs for finding all homologies execute in time essentially proportional to the square of the sequence length... » (page 5660). Again, these are lies.

LOCATING DENATURATION BUBBLES IN DNA

A very simple computer program was written, to locate regions of local melting in DNA, according to Azbel's theory, and the computational analysis of Gabarro-Arpa and Michel. See further down in the section on « epistemological work : the limitations of pattern analysis ».

OTHER BIOINFORMATICS WORK

GRAPHICAL CODING OF SEQUENCES

After the departure of Manolo Gouy, I lost enthusiasm for algorithmic work, and became more and more involved in visual perception. I was considering that the available tools for looking at sequences were providing straight answers to narrowly defined questions but did not provide enough space for serendipitous discoveries. Perhaps one should keep an open mind and develop graphic tools to visualize the sequences, look at them, detect potentially interesting features, and only then construct pertinent statistical tests and algorithms. Eduardo Mizraji, a biophysicist from Uruguay, having deep insights into all fields of theoretical biology, and with whom I had regular exchanges, was spending a month or two in my group, and he suggested a vectorial representation for nucleic acids. The sequence was represented by a trajectory in the plane, each nucleotide contributing to one step [32,33]. Importantly, the four vectors representing these steps did not add to zero and could be chosen with great flexibility. We had at least 12 different ways to look at a same sequence and often, only a few of them provided truly interesting shapes (for instance shapes that accurately reflected the intron exon subdivisions in genes-see the illustration in the end).

It turned out that similar vectorial representations had been already proposed, by Rosemarie Swanson in the domain of protein sequences [34], and by Hamori and Ruskin in the domain of nucleic acids [35], but we were not aware of these earlier contributions. A number of interesting features emerged in our work, among which the more or less "streamlined" character of biological sequences [33]. A bacterial sequence is subject to strong selective constraints, which seem to impose a strong homogeneity in local nucleotide composition. On the other hand, there are much larger fluctuations in the local composition of eucaryotic sequences. There was no follow-up to this work. I believe that it has a future, but the work should not be done superficially. It is necessary to have multiple looks at the same sequences, varying systematically the parameters of the graphical representations.

LOOKING FOR HIDDEN INFORMATION IN NUCLEIC ACID SEQUENCES : THE TOLERABLE NOISE PRINCIPLE

Let us take an RNA structure with stems composed of only G.C and A.U base-pairs. If you have an energy model which takes into account these pairs and forbids all other base pairs, it may succeed in predicting the RNA structure. But it will fail at predicting other structures that involve G.U pairing. Now the question is, if we do not know about these other RNA structures with G.U pairs, can we deduce something about G.U pairings just by analysing the RNA structures in which G.U pairs are not apparently used ? The counter-intuitive answer is: definitely yes, because if G.Us are possible, they generate negative constraints that can be detected.

Here is the procedure to evaluate these hidden possibilities. Let us incorporate in the energy model for predicting RNA structures, G.U pairs with an overestimated stability. For instance, we make in the model a single G.U pair as stable as four G.C pairs. Unless the RNA sequences under study are extremely idiosyncratic, when we will apply the folding algorithm, we will find erroneous structures, in which the numbers of G.U pairs are maximized. We get "parasitic structures". So, we can now assign to G.Us a lower stability, but as long as we get parasitic structure, this implies that we did overestimate the stability of G.U pairs. In this way, we can decrease the stability of GU.s until we reach a tolerable level of noise (not higher than the proportion of misfolded molecules in the test tube). So the stability of the G.U pair can be bounded on one side, by this argument. If we increase the number of sequences in our training set, we will get a more and more precise boundary, one which approaches better and better the true stability.

The true stability can be approached from the other side, using a set of structures in which G.U pairs are involved, and so cannot be underestimated. I believe (but have no formal proof, at present) that if we take a large enough set of sequences, the true stability can be approached from any side, to any desirable level of precision. This is due to the fact that the set of possible sequences is practically infinite, and that for random sequences, the number of potential folding schemes is enormous, generating a large number of alternative structures with free energies very close to the lowest free energy structure. So, there is "pressure" on both sides.

THE LIMITATIONS OF PATTERN ANALYSIS

There was, in the 1980's, a substantial body of work on "consensus sequences". It was believed that nucleic acid sequences could be analysed with word processors that would look for particular strings of letters, to which specific biological functions could be immediately assigned. On the other hand, people with a background in biophysics could legitimately discuss biological processes in terms of dynamical, physical interactions which minimized some thermodynamic variable. For instance, a biophysicist could try to predict the folding of a protein using a complex energy minimization algorithm. A bioinformatics expert could, on the other hand, try to predict the folding using statistical criteria - blind to the folding process- that would just state the chances that a given part of the protein sequence is or is not involved in a particular folding motif. If we believe that the ultimate truth is in the biophysical understanding of the folding process, what then is the validity of the bioinformatics approach through literal strings and consensus sequences ? I started discussing this epistemological question with Eduardo Mizraji, during one of his visits to Paris. We set up a gedanken experiment in which, having the physical model of a process which worked by energy minimization, we investigated how well we could predict its outcome by combining rules concerning the words in the sequences.

We were thus applying, to the domain of bioinformatics, a type of analysis that had been brilliantly developed by Minsky and Papert in their book on "Perceptrons" [36].

Having a pattern-analysis tool (an ancestor of modern neural networks), they constructed examples which the perceptrons could never solve. Therefore, against the popular belief that we will get better and better results by improving the methods, Papert and Minsky showed how some features inherent to the data could never be detected by some classes of algorithms. This was a very inspiring book. Impossibility statements are common in many domains of physics and mathematics (speed cannot travel faster than light, perpetual motion of the second kind is impossible, etc.), and the formulation of these impossibility principles are major landmarks in the history of science. In contrast, many people working in bioinformatics believe that they just need more computer memory, or faster computers, and they will be able to solve everything without changing radically their way of thinking.

So, in the spirit of Minsky and Papert, but at our much more modest level, we developed a test case for the confrontation of biophysics with bioinformatics. We chose as a test case the localization of melted regions in DNA. In a DNA sequence, the regions rich in A and T open more easily than the regions rich in G and C since G.C pairs are much more stable than A.T pairs. At a given temperature, a bare DNA molecule should present itself as an alternation of paired regions, rich in G and C, and unpaired regions, forming "denaturation bubbles", rich in A and T. There was a theory, by M.Y. Azbel [37] thanks to which the frontiers between paired and unpaired regions could be precisely determined, and the topic was under extensive experimental investigation in our institute, in the group headed by Claude Reiss. A practical computational approach for locating the frontiers was developed by Jaime Gabarro-Arpa and François Michel [38]). Using the simplified conditions of constant temperature and ionic strength, we devised a very simple algorithm (at most, 30 lines of code) to determine the frontiers, so we could deal with very large sets of sequences.

We thus generated large training sets of sequences, determined their partitioning into paired and unpaired regions, according to Azbel's model (it was a global energy-minimization model), and derived formal rules on the "words" present in the paired and the unpaired sequences. We then determined how well these rules succeeded on sequences outside the training set. With a large training set formed of random sequence, we found that the literal rules could achieve close to 99% success. On the other hand, if the training set was composed of sequences that had been derived from a common ancestor through extensive mutations, and considerable divergence, we found that the literal rules were contaminated with rules that reflected phylogenetic constraints. As we put it in the abstract [39], "(...) Thus, the global constraints imposed on sequences by a physical process may generate local patterns that are sufficient to predict, with a reasonable probability, the behaviour of the sequence. However, rather large sets of biological sequences are required to generate patterns free of illegitimate constraints. Furthermore, depending upon the initial sequence, the sets of sequences generated on a phylogenetic tree may be amenable or refractory to string analysis, while obeying identical physical constraints"

The manuscript was received by the reviewers with incredible hostility, as though it was directly threatening their job. They requested changes to insure that very few people could see what the article was about (it was about « the limitations of pattern analysis », my initial title, I think). In particular, one of them requested to remove the reference to Minsky and Papert's work. Nevertheless, I expected the article to be read and discussed, and become a classical paper in bioinformatics. I let you judge why things have turned differently.

THE EMBO MEETING ON "PATTERN ANALYSIS IN NUCLEIC ACID AND PROTEIN SEQUENCES" AT SAINT-AGNAN (1981)

I had the privilege to be the organizer of what was perhaps the first international bioinformatics meeting, under the banner of EMBO. There were 78 participants, with a strong representation from the USA and Canada (23 participants). Many participants were leaders or future leaders in their fields. The poster announcing the meeting is given at the end. Here is the text of the poster:

EUROPEAN MOLECULAR BIOLOGY ORGANISATION (E.M.B.O.)
WORKSHOP ON
PATTERN ANALYSIS IN NUCLEIC ACID AND PROTEIN SEQUENCES

Saint-Agnan, Bourgogne (France) October 27th- 30th, 1981.

THE LESSONS. One aim of the workshop is to evaluate critically the work that has been done or can be done in the exploitation of sequence data. EVOLUTION. Construction of phyletic trees, search for homologies, guesses on ancestral sequences. STRUCTURE. Prediction of secondary structures in proteins and nucleic acids, of folding domains in proteins, of melting domains in nucleic acids. PATTERNS. Search for repeats, palindromes, periodicities, cleavage sites, recognition signals. CONSTRAINTS. How the genetic code and other factors may constrain the sequences. Sequence reconstruction from fragments. Design of optimal sequences for synthesis.

THE TOOLS. The second aim of the workshop is to explore all possible ways of going further in the exploitation of the sequence data. EXISTING TOOLS. Efficient algorithms. Editing and presenting the data. Graphic analysis. Statistics. ARTIFICIAL INTELLIGENCE. The methods of pattern analysis in other fields: image enhancement (e.g., in electron microscopy) Fourier analysis (crystallography), the methods of structural linguistics, adaptive filters and adaptive memories (perception).

Inquiries and applications: write before April 30th to: Jacques Ninio/ IRBM, Tour 43/ 2 Place Jussieu / 75251 Paris cedex 05 / France

----

The meeting took place in a vacation camp far from any village, so all the participants stayed together for the whole duration of the meeting. The format was that of the early EMBO meetings, with no talk scheduled in advance. The chairpersons were chosen the day before their session, and they organized their session with all participants willing to intervene. After giving an introduction to their field stressing the points of agreement and the points in debate, the other contributors would develop their personal themes. The program of the four days is given below, with the names of the chairpersons between brackets:

Tuesday 27th
Morning: Protein architecture and folding (Jane Richardson and Michael Sternberg)
Phyletic trees (Michael Hendy)
Afternoon: Data banks and computer tools (Carolyn Tolstochev)
Sequence determinations (Gary Ruvkun)
After dinner: Computer graphics (Robert Langridge and Richard Feldman)

Wednesday 28th
Morning: Sequences and genetics (Douglas Brutlag)
Protein patterns and predictions (Shoshana Wodak)
Afternoon: Excursion in the wine caves of Burgundy and in Beaune's hospice
After dinner: Fourier analysis, statistics (David Sankoff)

Thursday 29th
Morning: codon usage (Jacques Ninio)
Protein evolution (David Penny)
Afternoon: Patterns in nucleic acids (Edward Trifonov)
Open themes: Jean-Luc Darlix
After dinner: Artificial intelligence (Peter Friedland, Jean Sallantin)

Friday 30th
Morning: Protein design (Athel Cornish-Bowden)
Nucleic acid evolution (Robert Cedergren)
Afternoon: RNA secondary structure prediction (Ann Jacobson)
Evaluation of the meeting (Pierre Oudet)
After eight: Banquet

REFERENCES

[1] Dumas, J-P. and Ninio, J. (1982) Efficient algorithms for folding and comparing nucleic acid sequences. Nucleic Acids Res. 10, 197-206.

[2] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

[3] D.E. Rumelhart, Hinton, G.E., and Williams, R.J. (1986) Learning representations by back-propagating errors. Nature 323, 553-556.

[4] Lake, J.A., and Beeman, W. W. (1968) On the conformation of yeast transfer RNA. Journal of Molecular Biology 31, 115-125.

[5] Lake, J. A. (1967) An iterative method methof of slit-correcting small-angle X-ray data. Acta Crystallographica 23, 191-194

[6] Ninio, J. (1971) Etude de la structure de l'ARN de transfert par diffusion centrale des rayons X, et de ses implications biologiques. Thèse d'Etat, Université Paris 7.

[7] Pavlov, M. Y. and Fedorov, B. A. (1983) Improved technique for calculating X-ray scattering intensity of biopolymers in solution : evaluation of the form, volume, and surface of a particle. Biopolymers 22 1507-1522.

[8] Koch, M. H. J., Vachette, P., and Svergun, D. I. (2003) Small-angle scattering : a view on the properties, structures and structural changes of biological macromolecules in solution. Quarterly reviews of biophysics 36, 147-227.

[9] Ninio, J., Luzzati, V. and Yaniv, M. (1972) Comparative small-angle X-ray scattering studies on unacylated, acylated and cross-linked *Escherichia coli* transfer RNA Val. 1. J. Mol. Biol. 71, 217-229.

[10] Pipas, J. M., and McMahon, J. E. (1975) Method for predicting RNA secondary structure. Proc. Nat. Acad. Sci. USA 72, 2017-2021.
[11] Ninio, J. (1979) Prediction of pairing schemes in RNA molecules. Loop contributions and energy of wobble and non-wobble pairs. Biochimie 61, 1133-1150.

[12] Leontis, N. B., and Westhof, E. (1998) Conserved geometrical base-pairing patterns in RNA. Quarterly review of biophysics 31, 399-455.

[13] Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. Journal of Molecular Biology 288, 011-940.

[14] Papanicolaou, C., Gouy, M. and Ninio, J. (1984) An energy model that predicts the correct folding of both the tRNA and the 5S RNA molecules. Nucleic Acids Res. 12, 31-44.

[15] Gouy, M., Marlière, P., Papanicolaou, C. et Ninio, J. (1985) Prediction des structures secondaires dans les acides nucléïques: aspects algorithmiques et physiques. Biochimie 67, 523-531.

[16] Woese, C.R., Gutell, R., Gupta, R., and Noller, H.F. (1983) Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. Microbiology Reviews 47, 621-669

[17] Traub, W., and Sussman, J. L. (1982) Adenine-guanine base pairing in ribosomal RNA Nucleic Acids Research 10, 2701-2708.

[18] Abou-ela, F., Koh, D., Tinoco, I. Jr, and Martin, F.H. (1985) Base-base mismatches. Thermodynamics of double helix formation for dCA3XA3G + dCT3YT3G (X, Y = A, C, G, T). Nucleic Acids Research 13, 4811-4824.

[19] Cech, T. R., Tanner, N. K., Tinoco, I. Jr, Weir, B. R., Zuker, M., and Perlman, P.S. (1983) Secondary structure of the tetrahymena ribosomal RNA intervening sequence : structural homology with fungal mitochondrial intervening sequences. Proc. Nat. Acad. Sci. USA 80, 3903-3907.

[20] Haralick, R. M., and Elliott, G. L. (1980) Increasing tree search efficiency for constraint satisfaction problems. Artificial intelligence 14, 263 – 313.

[21] Dumas, J.-P. (1981) Etude de la modélisation des acides ribonucléiques, thèse de l'université Pierre et Marie Curie, Paris 6. Dumas states in the introduction to his thesis that he had no part in the fast comparison algorithm, his work being exclusively in the secondary structure prediction domain.

[22] Nussinov, R., and Jacobson, A.B. (1980) Proc. Nat. Acad. Sci. USA 77, 6309-6313.

[23] Zuker, M., and Stiegler, P. (1981) Nucleic Acids Research 9, 133-148.

[24] Williams, A. L., Jr, and Tinoco, I. Jr (1986) A dynamic programming algorithm for finding alternative RNA secondary structures. Nucleic acids research 14, 299-315.

[25] Marlière, P. (1983) Computer building and folding of fictitious tRNA sequences. Biochimie 65, 267-273.

[26] Lee, Y.-H., Dsouza, L.M., and Fox, G.E. (1997) Equally parsimonious pathways through an RNA sequence space are not equally likely. Journal of Molecular Evolution 45, 278-284.

[27] Fontana, W., and Schuster, P. (1998) Shaping space : the possible and the attainable in RNA genotype-phenotype mapping. Journal of Theoretical Biology 194, 491-515.

[28] Ninio, J. (1991) Transient mutators: a semi-quantitative analysis of the influence of translation and transcription errors on mutation rates. Genetics 129, 957-962.

[29] Ninio, J. (1996) Gene conversion as a focusing mechanism for correlated mutations: a hypothesis. Molecular and General Genetics 251, 503-508.

[30] Wilbur, W.J. and Lipman, D.J. (1983) Rapid similarity searches of nucleic acid and protein data banks Proc. Nat. Acad. Sci USA 80, 726-730.

[31] Karlin, S., Ghandour, G., Ost, F., Tavare, S., and Korn, L.J. (1983) New approaches for computer analysis of nucleic acid sequences. Proc. Nat. Acad. Sci USA 80, 5660-5664.

[32] Mizraji, E. and Ninio, J. (1985) Graphical coding of nucleic acid sequences. Biochimie 67, 445-448.

[33] Ninio, J. et Mizraji, E. (1995) Perceptible features in graphical representations of nucleic acid sequences. In *Visualizing biological information* (C.A. Pickover, ed.) World Scientific, Singapore, PP. 33-42.

[34] Swanson, R. (1984) A vector representation for amino acid sequences. Bulletin of Mathematical Biology 46, 623-639

[35] Hamori, E., and Ruskin, J. (1983) H curves, a novel method of representation of nucleotide series, especially suited for long DNA sequences. Journal of Biological Chemistry 258, 1318-1327.

[36] Minsky, M., and Papert, S. (1969) Perceptrons : an introduction to computational geometry. The MIT Press, Cambridge, MA.

[37] Azbel, M. Y. (1980) DNA sequencing and helix-coil transition. I. Theory of DNA melting. Biopolymers 19, 61-80

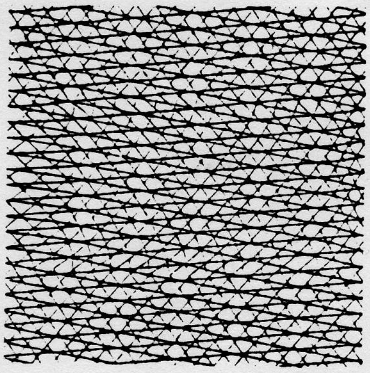
[38] Gabarro-Arpa, J., and Michel, F. (1982) The hierarchical approach to the DNA stability problem. I. - Patterns in non-equilibrium denaturation and renaturation.

[39] Ninio, J. and Mizraji, E. (1989). String analysis and energy minimization in the partition of DNA sequences. J. Mol. Biol. 207, 585-596.

Other articles relevant to bio-informatics

Ninio, J. (1983) L'explosion des séquences: les années folles 1980-1990. Biochemical Systematics and Ecology 11, 305-313.
A tongue-in-the-cheek essay on the future developments of sequence analyses

Ninio, J. (2000) Illusory defects and mismatches: why must DNA repair always be (slightly) error prone ? BioEssays 22, 396-401.
This article draws the attention to a particular class of negative constraints in DNA sequences, that can be revealed in oligonucleotide statistics.



EUROPEAN MOLECULAR BIOLOGY ORGANISATION (E.M.B.O.)

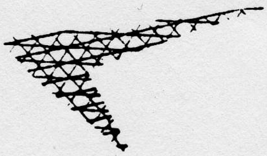
WORKSHOP ON

PATTERN ANALYSIS IN NUCLEIC ACID AND PROTEIN SEQUENCES

Saint-Agnan, Bourgogne (France) October 27th -30th, 1981

THE LESSONS . One aim of the workshop is to evaluate critically the work that has been done or can be done in the exploitation of sequence data. *EVOLUTION*. Construction of phyletic trees, search for homologies,guesses on ancestral sequences.*STRUCTURE*. Prediction of secondary structures in proteins and nucleic acids, of folding domains in proteins,of melting domains in nucleic acids. *PATTERNS*.Search for repeats,palindromes,periodicities,cleavage sites,recognition signals.*CONSTRAINTS*.How the genetic code and other factors may constrain the sequences.Sequence reconstruction from fragments.Design of optimal sequences for synthesis.

THE TOOLS.The second aim of the workshop is to explore all possible ways of going further in the exploitation of the sequence data. *EXISTING TOOLS*.Efficient algorithms.Editing and presenting the data.Graphic analysis.Statistics. *ARTIFICIAL INTELLIGENCE*.The methods of pattern analysis in other fields : image enhancement (e.g.,in electron microscopy) ,Fourier analysis (crystallography), the methods of structural linguistics , adaptive filters and adaptive memories (perception) .



Inquiries and applications : write, before April 30th to :

Jacques Ninio / IRBM, Tour 43 / 2 Place Jussieu /
75251 Paris cedex 05 / France

same representation, the *E. coli* 16S RNA appears to be more homogeneous. It is not unlikely that in bacteria (or more generally, in organisms with small genomes), there are general selective constraints on DNA sequences which tend to promote statistical homogeneity all along the genes.

The last example is that of human α-hemoglobin (Fig. 5). The separation in introns and exons is obvious. The relationship between gene and pseudogene is less obvious. The positioning of the intron-exon junction suggests that the frontiers may have moved somewhat, due to nibbling of introns by adjoining exons. The introns of β-hemoglobin, shown in reference 3, seem unrelated to those of α-hemoglobin.

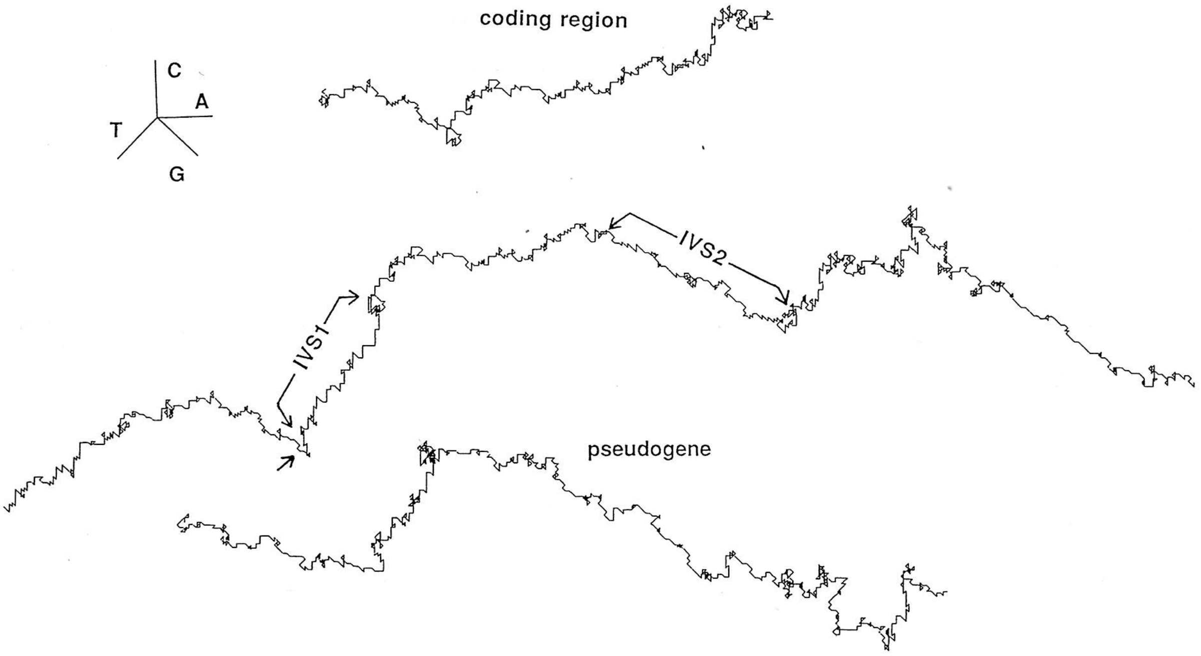


Fig. 5. Human α-hemoglobin. The sequence in the middle is that of a human α-hemoglobin gene.[19] The introns are labeled IVS1 and IVS2, and the complete coding region is shown on top. The bottom sequence is that of a pseudogene for human α-hemoglobin.[20]

One conspicuous property of vectorial representations is that they present compressed or expanded regions. Roughly, a compressed region is one in which there is a high turnover of the four nucleotides. In extended regions, there is a bias against one or more nucleotides. In all cases examined so far, the compressed/extended criterion failed to lead to biologically important insights. On the other hand, the dissection of the sequence in regions of constant local direction often reflects known biological features.

Three different correspondences between vectors and nucleotides were used for the three examples of Figs. 1–5. Twelve different codings are possible,