BIOINFORMATICS

Vol. 16 no. 1 2000
Pages 24–33

On the trail of protein sequences

Russell F. Doolittle

Center for Molecular Genetics, University of California, San Diego, La Jolla, CA 92093-0634, USA; E-mail: rdoolittle@ucsd.edu

Introduction

I was somewhat taken aback when asked to write an article for a History issue of *Bioinformatics*, because not by any stretch of the imagination am I a 'bioinformaticist'. I have no formal training in computer or information science. By education, I am a biochemist whose early experience was in the area of proteins. Bioinformatics was not a term that existed when I began my scientific career.

My introduction to computers came about from an interest in biochemical evolution, a subject that first fascinated me many years ago when I was a graduate student. The laboratory in which I did my graduate training was working on blood proteins—especially those involved in blood coagulation—and a number of chance factors led me to inquire how this quite complicated process could ever have evolved. Blood clotting in humans was known to depend on the coordinate interplay of a dozen or more protein factors. This was a period when the notion of one gene–one polypeptide chain was beginning to be generally accepted, and it seemed unlikely to me that the entire melange could have evolved in one fell swoop. Rather, there must have been a series of gene duplications involving these clotting factors, just as had been recently suggested for some of the chains of hemoglobin. I was an early advocate of the 'all new proteins from old proteins' school of thought.

The question arose, if one knew the amino acid sequences of all these clotting proteins, could the order of the duplicative events be reconstructed? As it happened, the question was moot, because none of their sequences was known at the time, and determining even one of them would have been an arduous undertaking. Accordingly, I took a different tack, concentrating on a search to find the phylogenetic distribution of the clotting factors. In particular, I sought out the most primitive creatures which exhibit the thrombin-catalyzed conversion of fibrinogen to fibrin. In the end, I found that all vertebrates, even jawless fish like the lamprey, had this ability (Doolittle *et al.*, 1962).

Determining amino acid sequences

The conversion of fibrinogen to fibrin is initiated by the release of small peptides from the parent fibrinogen molecule. These peptides—called fibrinopeptides—tend to be extremely variable in sequence, and I thought they would be good markers to follow the course of species divergences. In 1964, I went to Sweden on an NIH postdoctoral fellowship to learn the art of the Edman degradation in the laboratory of Birger Blomback. We sequenced numerous fibrinopeptides (Doolittle and Blomback, 1964) and showed that sequence comparisons were good reflections not only of the newly emergent genetic code but also of relationships inferred on the basis of the fossil record (Figure 1). I continued the project

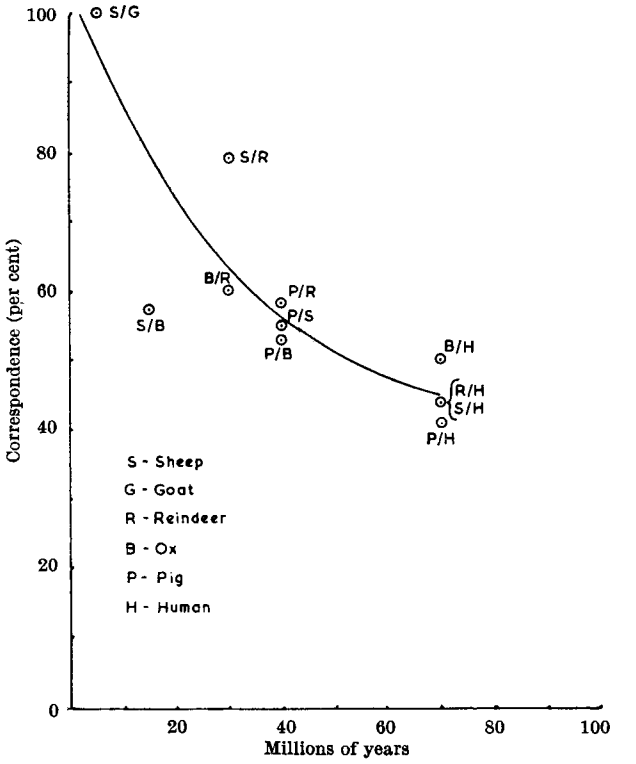

Fig. 4. Amino-acid sequence correspondence between fibrinopeptides of various pairs of mammals plotted against time since last common ancestor. Percentage correspondence based on number of identical amino-acids in the same sequential position in both fibrinopeptides *A* and *B* (Table 1)

Fig. 1. Legend and figure reprinted from Doolittle and Blomback (1964).

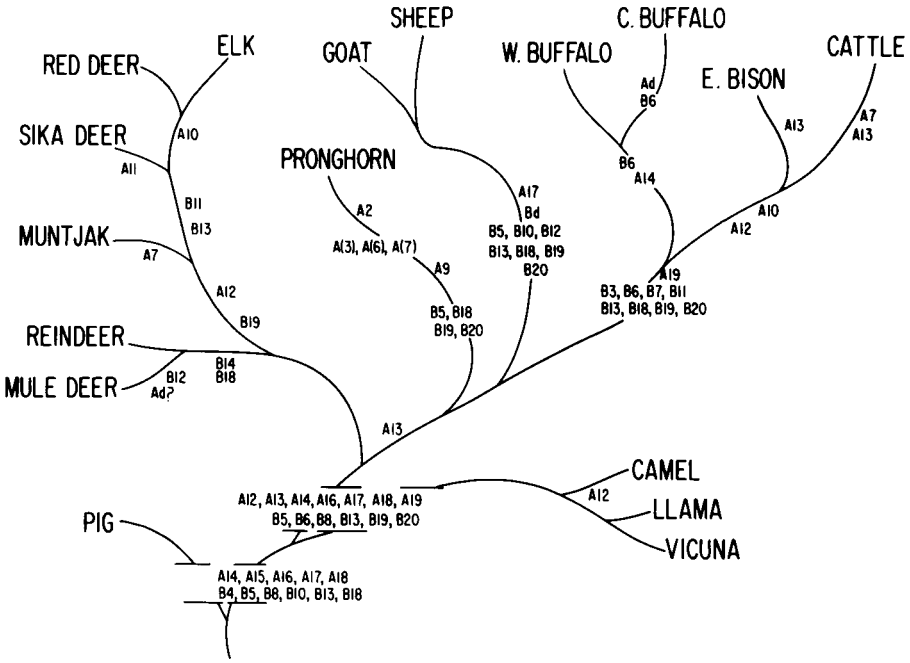

FIG. 3. Classical relationship of 18 artiodactyls and a mutational scheme consistent with the present day fibrinopeptide amino acid sequences. Letter-number designation indicates position in fibrinopeptide A or B in which an amino acid replacement has occurred on that branch of the phylogenetic tree. Deletions are designated by "d." On the branch leading to the pronghorn antelope three of the changes are noted in parentheses. At least three such amino acid replacements have occurred in this region of the fibrinopeptide A, but their exact positions are still uncertain.

Fig. 2. Figure and legend reprinted from Mross and Doolittle (1967).

when I moved to San Diego (e.g. Mross and Doolittle, 1967), as did Blomback in Sweden (Blomback *et al.*, 1966).

Because of the high variability of the fibrinopeptides, there were numerous amino acid replacements to assess (Figure 2). In general, the strategy was to assume that the paleontologists had got it right, and we simply incorporated the observed changes in the presumed tree (Figure 2). Nonetheless, there were cases when the simplest ordering of events seemed out of line with the fossil record (Doolittle and Blomback, 1964). Indeed, I had an interesting exchange of letters with George Gaylord Simpson about the fossil record of artiodactyls. Simpson was skeptical of the fibrinopeptide data, and I was anxious to develop a method that could cluster the taxa on the basis of the sequence data alone. One of my students, Susan Tideman, managed to write a computer program that constructed a matrix of the minimum number of (DNA) base changes needed to explain the amino acid replacements (Figure 3), but we were stymied as to what to do next until Fitch and Margoliash (1967) and Dayhoff and Eck (1968) published their elegant procedures for tree building.

In truth, I had been aware of computer power for some time. In 1954, when I got out of the Army I found a job as a computist (a kind of engineer assistant) in the Research and Development section of a large aircraft company. Mostly the job was to use an electric (not electronic) calculator to process large amounts of data collected at a jet engine test facility. The most sophisticated device routinely available to us was a simple punched-card calculator. But one day we were all gathered together and given a lecture on a real digital computer (IBM 703). I hadn't the faintest idea how it worked, but its arithmetic capabilities were awesome. It made a big impression on me, and a decade later, with known sequences beginning to accumulate, the benefits of using a computer to manage the data seemed obvious.

When I arrived in San Diego in 1964, I sought out the fledgling campus computer center and met with a 'consultant'. The experience convinced me that either computer people were going to have to learn biology, or I'd have to learn about computers. There was nothing I could do to influence the former, so the latter course was

	Pig	Camel	Vicuna	Elk	Sika deer	Muntjak	Reindeer	Mule deer	Pronghorn	Sheep	Persian gazelle	Water buffalo	Cape buffalo	Eur. bison	Ox	Human
Pig	-															
Camel	14		-													
Vicuna etc.	13	1	-													
Elk	18	15	15	-												
Sika deer	17	14	14	1	-											
Muntjak	16	14	14	5	4	-										
Reindeer	17	14	15	7	6	4	-									
Mule deer	17	12	13	8	7	5	1	-								
Pronghorn	23	18	19	11	12	13	13	14	-							
Sheep etc.	19	16	17	15	14	12	11	9	16	-						
Persian gazelle	18	15	16	11	10	10	10	10	15	10	-					
Water buffalo	19	14	15	16	15	14	13	13	18	17	15	-				
Cape buffalo	14	13	13	14	13	11	12	12	18	17	12	1	-			
Eur. bison	18	16	16	13	14	13	14	14	21	20	16	5	3	-		
Ox	20	18	18	15	16	15	16	16	22	22	18	7	4	2	-	
Human	23	18	19	22	23	20	20	20	26	21	20	20	16	19	21	-

Fig. 8. Computer tabulation of minimum number of (DNA) base changes necessary to explain amino acid replacements in fibrinopeptides A and B among various artiodactyls and human.

Fig. 3. Table prepared by Susan Tideman in the author's laboratory in 1967; it eventually appeared in Doolittle (1970).

the only option. I enrolled in a course in diagnostic Fortran and learned a few simple things. My progress was slow. I was an experimentalist, and the laboratory took up most of my energy. Happily, I had a sub-teenager son who was keen on math and computing and was soon writing simple programs for me. Also, Walter Fitch sent me the code for some of his programs. Slowly, I was getting the hang of it. The goals throughout this period were to categorize protein sequences and to find relationships between them.

In 1972, I was invited to write an article on protein evolution for a new volume of *The Proteins*, edited by Hans Neurath and Robert Hill. This was a wonderful opportunity to sort out all my thoughts on the subject, and I threw myself into the project fully. I incorporated a number of computer aspects and included some newly identified homologies. Unhappily, there was a 6-year delay between submission and publication (Doolittle, 1979), during which time others had come to many of the same conclusions. Some of the predictions of homology made in the article have only recently been borne out by X-ray structures (Figure 4).

During this period, for reasons of technique and strategy, most people who were sequencing proteins were very much aware not only of their own experimental results, but also of those being reported by others. As a result, unexpectedly homologous sequences were being found even without the aid of computers. For example, lactalbumin was found to resemble lysozyme (Brew *et al.*, 1967), and haptoglobin looked like a dead serine protease (Barnett *et al.*, 1972). Increasingly, gene duplication was being recognized as the major force in generating both new and longer proteins.

Most workers in the sequence field spent their evenings poring over their sequences trying to make sense of them. I well remember Jim Brown visiting me in 1973 and telling me the story of lining up paper strips with his sequences typed on them on his living room floor and discovering that serum albumin had a triplicated structure.

Sequence searching by computer in the 1970s was very much the monopoly of Margaret Dayhoff and her colleagues at the National Biomedical Research Foundation (NBRF). It was just a little annoying to experimentalists to submit one's results to that center and then have to wait for increasingly long periods before the issuing of the fully compiled data. There were two reasons for the delays. For one, the staff at the

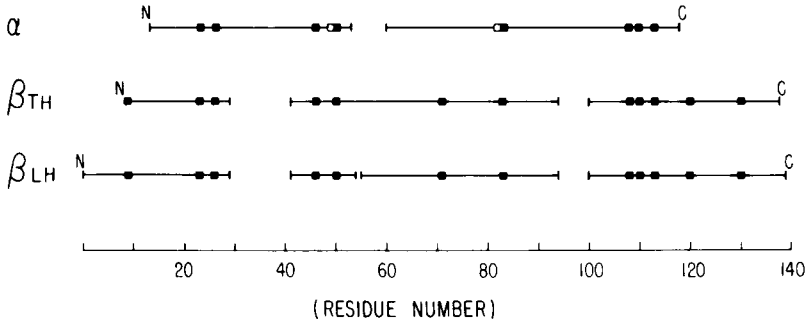

Fig. 33 **Comparison of the identical α-chain subunits of bovine thyrotropin (TH) and luteinizing hormone (LH) with their β-chains as to the location of cysteine residues (all of which are involved in disulfide bridges in the native molecules). The solid squares indicate aligned cysteines; two unmatched cysteines in the α-chain are denoted by hollow squares. Computer comparisons (Dayhoff, 1972) of the sequences have been unable to detect common ancestry between the α-chains and the β-chains.**

Fig. 4. Figure and legend reprinted from Doolittle (1979). X-ray crystallography has since shown that the α and β chains of gonadotropins are indeed homologous (Wu *et al.*, 1994).

NBRF were meticulously annotating the data, adding many interesting features. They were also looking for interesting connections themselves, often publishing their own findings in the *Atlas* itself. In those days, the *Atlas of Protein Sequence and Structure*, in its periodic appearances, was a collection heavily biased by cytochromes c, immunoglobulin light chains, hemoglobins and fibrinopeptides, a natural consequence of protein biochemists attacking peptides and small proteins first. As such, it was hardly representative of proteins in general and was not particularly useful for generalizing about structural features. The thought occurred to me that we ought to start our own sequence collection.

Let me hasten to add that, my complaints aside, I have always been a great admirer and staunch supporter of the efforts of the NBRF. I personally learned more from the sundry volumes of the *Atlas of Protein Sequence and Structure* (beginning with Eck and Dayhoff, 1965) than any other source. The Minimum Mutation Matrix, in my opinion, stands alongside the Needleman–Wunsch (1970) algorithm as one the most elegant contributions in the entire field of sequence analysis. But this reverence not withstanding, I feel my complaints were justified.

Back in the 1960s and 1970s, computers in academic institutions tended to be located in campus computer centers, where they were mostly used as number crunchers for the physical sciences, on the one hand, and for business and accounting, on the other. The 80-column punched card was the main medium, and typically stacks of cards were left off at the computer center in the afternoon with the hope of retrieving output the next morning. After a while, some science departments acquired their own computers, and then, gradually, some individual research groups were able to get their own (but only with permission of the campus computer committee, which worried about losing support for the central facility). In 1976 I managed to acquire a DEC PDP-11, and we began logging our own sequences.

The DNA revolution

I'm a protein person, and I have steadfastly avoided DNA (or RNA) unless the sequences were translated into protein. No promoters or enhancers or other non-protein entities. Of course, today, the vast majority of protein sequences are known from DNA sequencing, without which the data banks would be only modestly, larger, but more manageable, than they were in the middle 1970s.

In 1978, one of my colleagues, Ted Friedmann, returned from a sabbatical leave with Fred Sanger in Cambridge, where he had learned the then 'miraculous art' of DNA sequencing. Ted had been sequencing the oncogenic DNA virus called polyoma. Gernot Walter, who was then at the Salk Institute, had gotten hold of some DNA sequence information from another oncogenic virus, SV40, and the two of them, having learned that we were comparing sequences by computer, came to visit me to see if the computer could identify any resemblances, something they had been unable to ascertain by eye. The regions of these viruses for which they had sequences were the 'small tumor antigen'. In fact, after translation, the sequences proved to be 28% identical. Was that significant, they wondered? It was the same as the resemblance of myoglobin and hemoglobin chains, I pointed out. Statistics aside, they found that convincing. We also searched the sequences against a newly purchased

NBRF tape and our own growing collection; nothing else came close to being similar. We wrote a note about it (Friedmann *et al.*, 1978).

More experimental ties

Even before Ted Friedmann and Gernot Walter visited, the computer had become an integral part of our research. Our laboratory was involved in both sequencing proteins and chemically synthesizing peptides. In the first case, we were deep into a project sequencing the three homologous polypeptide chains of human fibrinogen, and we had been using the computer both to align the sequences, on the one hand, and to identify the α-helical stretches that form coiled coils in the molecule by the Chou and Fasman (1974) approach, on the other (Doolittle *et al.*, 1978).

The synthetic peptides were being used in a variety of settings, among which was their attachment to carriers for raising antibodies for use as probes in our structural studies. In follow-up discussions to the polyoma–SV40 comparison, Gernot Walter wondered how he might be able to distinguish the newly discovered splicing products of the SV40 large and small T antigens, which shared a common amino-terminus but had different carboxyl termini. Could antibodies directed to synthetic peptides make the distinction? Gernot went to work in our laboratory, and with the help of one of my graduate students, Andy Laudano, made a series of peptides. Antibodies were raised to appropriate peptide-carrier conjugates, and, back in his own lab, Gernot used them to immunoprecipitate radiolabeled virus proteins. The results were spectacular (Walter *et al.*, 1980). The novel aspect of this work was the revelation that antibodies could be generated for proteins which had never been isolated and were known only from their DNA sequences.

Hydropathy plots

During the period 1977–1987, I had the benefit of having my laboratory right around the corner from Jack Kyte. One evening on my way out of the building, I showed him a plot of a similarity profile I had just constructed for the homologous sequences of the β and γ chains of fibrinogen. The display used a simple moving window that showed the number of identities in each overlapping 20-residue segment. Jack stared at it for a moment, and then suggested what I *really* ought to do was write a program that could display the hydrophobic character of a protein. I think the program was ready the next day (one of my sons helped me de-bug it that evening). Jack and I then spent the next 6 months in a daily discussion arguing over a hydropathy scale for the 20 amino acids.

Indeed, the use of the word hydropathy emerged from those discussions. We were searching for a term that would encompass both hydrophobicity and hydrophilicity, and hydropathy won out because of its appropriate roots. It was based on a ‘feeling for’ (-pathy) water (hydro-). The usage was vexing to some, and one well known protein scientist wrote a letter to the *Journal of Molecular Biology* complaining that ‘hydropathy was a 19th century water cure for unknown ailments’ and protein science should not be corrupted by such inept terminology. But word usages can change over the centuries, even in England, and hydropathy has obviously stuck.

Terminology aside, the program proved especially useful in two realms. First, we employed it to great advantage in choosing sequences for synthesizing peptides for raising antibodies to proteins known only from their sequences. Second, it was very effective in identifying membrane-spanning sequences.

While we were getting the paper together, it was pointed out to us that George Rose (1978) had anticipated us with a program that predicted turns on the basis of a hydrophobicity measure. In fact, his smoothing procedure was really more sophisticated than ours. Additionally, the quite similar method of Hopp and Woods (1981) appeared, and subsequently we found that the same idea had been depicted in a paper reporting the sequence of flu virus hemagglutinin (Both and Sleigh, 1980). Nothing like an idea whose time has come! Nonetheless, I have always looked back on our own paper with great fondness. I frequently quip that it is one of the ‘most cited, least read papers’ in all of protein chemistry.

Funding

The year 1979 marked a turning point in our laboratory. After many years of effort, we had completed our amino acid sequence work on human fibrinogen (Doolittle *et al.*, 1979). We needed a change, and it seemed to me reasonable to exploit our recent success with computer analysis of sequences, modest as it may have been.

In 1978 I had heard a lecture by H.G.Wittmann on his herculean studies of ribosomal proteins, during which he made the offhand remark that none of the more than 40 proteins they had characterized were homologous with each other. As I stared at the slides he was showing, I couldn’t help but think that some of those sequences certainly *looked* homologous. Inspired both by the challenge and the availability of this large data set of small proteins, I set to work with Neal Woodbury, an undergraduate who was guiding me through this early stage of computer independence, and Rodney Jue, a beginning graduate student, characterizing the ribosomal protein sequences as a test for our programs and abilities. I was convinced that these foundational proteins, likely heirlooms from the earliest stages of life, ought to be related to each other and would also contain internal duplications reflecting their primitive beginnings. We all know that such preconceived

notions are dangerous in science, although they're usually easier to spot in others.

The project was worthwhile on several fronts, and we learned a good deal about computing and sequence comparison both, even though we were operating under severe constraints. The PDP-11 had a hard disk with a maximum capacity of only 100 kilobytes, much of which was occupied by a mini-Unix operating system. Neal Woodbury wrote a variety of programs (Figure 5), the most important of which was a simple search routine patterned on an x/y moving-window approach used by Fitch (1966) for alignment purposes. Neal also wrote a version of the Needleman–Wunsch alignment scheme, which, because of the limits of the PDP-11, was limited to sequences of 90 residues or less.

Eventually, we submitted a paper (Jue *et al.*, 1980) to the *Journal of Molecular Evolution* that reflected my preconceived notions about ribosomal proteins. The original version of the paper was soundly criticized by a knowledgeable reviewer, who ran several of our comparisons through his or her own computer system and found them to lack statistical validity. We heeded some of that advice and removed some of the weaker claims. Nevertheless, we still contended that we had shown some of the ribosomal proteins to be related to others, and in several cases stuck with the idea of internal repeats. In the end, Emile Zuckerkandl, the most interactive of journal editors, accepted the paper on the advice of another reviewer who wrote: '…the potential interest seems to outweigh the risk that the results are illusory'.

In retrospect, some of the results were illusory. In one case, in a comparison we had labeled a 'case of certain homology', it turned out that one of the proteins had been mis-sequenced, apparently having been contaminated with one or more peptides from the other protein. The sequence analyst is always at the mercy of the experimentalist! In some other cases, the support for internal duplication was thin even by the innocent criteria we had imposed. Some of the relationships may yet turn out to be valid, however, and I am hoping that more three-dimensional structures of these proteins will appear soon and settle the matter.

But the paper was better than the dubious conclusions may imply. One of the programs that Neal had written (but not listed in the table shown in Figure 5) was a three- dimensional version of the Needleman–Wunsch algorithm. Programming this was a *tour de force*, even though the constraints of our computer limited comparisons to 20-residue lengths. Five years later a comparable three-dimensional Needleman–Wunsch was reported for more capable computers (Murata *et al.*, 1985).

Once the paper was submitted, I tried to embody the results in part of a grant application to the National Science Foundation (NSF) on the subject of distant relationships. Obviously we needed access to a larger computer, and the grant asked for funds to buy a link to the Chemistry Department's new VAX. I worked very hard on the application, and when it was submitted I felt confident it would be funded. Six months later, however, the bad news arrived. Although most of the eight anonymous reviewers ranked the proposal 'excellent' or 'very good', two were not so generous. One, with a rating of 'fair', was a mostly well-reasoned critique, the thrust of which was that I seemed unaware of the work of experts in the field of sequence comparison, although it was unclear to me how I was supposed to know about the work of Smith and Waterman 'in preparation'. The review also noted that, with regard to a particular approach, '… Smith had already developed a program for doing this and concluded that it was impractical'.

The review that really sunk the proposal, however, ranked it 'poor' and unabashedly stated that 'we already do all this at the *Atlas of Protein Sequence and Structure'*.

A few weeks after the arrival of the bad news from the NSF, I received a brief letter from Temple Smith and Mike Waterman:

> Dear Dr. Doolittle:
>
> Your recent paper in *J. Mol. Evol.* suggests you may not be familiar with some of the more recent sequence comparative metrics, thus we have enclosed for your possible interest our most recent work.

The timing was awful, and I let loose with both barrels. It was obvious that our *JME* paper had been submitted well before one of the two provided papers had appeared, and the other had been published in *Advances in Mathematics*! I pointed out that by coincidence the same points had been made by a reviewer of our recently denied NSF application. In response, both immediately wrote back protesting that neither had been involved in reviewing the grant. Indeed, in reading this correspondence some 19 years later, I see in that second letter the same gentle Mike Waterman with whom I would later become good friends, almost apologetic about publishing where biologists were unlikely to roam, and offering to get together to explore common ground.

The funding problem was perennial. A year or two later, I was invited to be a part of a departmental program project application aimed at getting support for a variety of computer applications. My participation would have gained me that access to the Chemistry Department's VAX computer that I so longed for. During my presentation to the site-visit committee (which included one of the contributors to this special issue of *Bioinformatics*), I made the mistake of referring to my computer efforts as 'pretty much of a hobby' (I was overly proud of being an experimentalist). In any case, the committee, perhaps

Table 1. Approximate time requirements for several minicomputer programs described in this article[a]

Program	Function	Time required for consideration of two sequences 1-90 Residues	100-999 Residues
SEARCH	Find x identities in a sliding segment y residues in length	< 1 min	~1 min
ALIGN	Align two sequences using sliding segment approach		
	Unitary Matrix	3-5 min	10-30 min
	Homology Matrix	5-15 min	0.5-3 h
NEEWU	Use Needleman-Wunsch algorithm to determine optimum alignment	5-10 min	0.5-3 h
NWJUM	Statistical verification of Needleman-Wunsch alignments	2-4 h	

[a] Times are for a PDP 11/04 based system as described in text

Fig. 5. Table from appendix of Jue *et al.* (1980) reflecting constraints of a small computer.

anxious that I not compromise my amateur standing, approved the overall project but deleted my section. The trend continued, and by and large most of my computer work over the years has been bootlegged off of grants awarded for laboratory work. In the end, it was all for the best.

Matchmaking

It is probably fair to say that the bulk of my notoriety in the area of what was destined to become 'bioinformatics', was gained during the 1980s from searching newly determined sequences against our own database. I have recently described some of my favorite 'hits' elsewhere (Doolittle, 1997), and I will resist the urge to list them all again. Instead, I'll only say that we began with some clumsy home-made programs (Figure 5), a tiny, slow computer and a winning strategy. The strategy was to type new sequences into our computer as fast as they appeared and immediately search them against all other known sequences. Moreover, we eschewed entering obviously redundant sequences like hemoglobins, immunoglobulins and cytochromes c.

The sequence-entering was slavish work, the brunt of which was borne by my secretary, Karen Anderson, and my younger son, Will (although I did my share). A routine was established whereby every sequence was verified, searched, its composition, secondary structure and hydropathy plot depicted, all these vitals, including a photocopy of the primary source, whether it be a journal or someone's handwritten scrawl, being stored in a manila folder and filed as hardcopy as well as electronically. We kept a complete citation log, but we skipped the detailed annotation and curation that was bogging down the *Atlas of Protein Sequence and Structure*. Because the collection was an obvious extension of the NBRF Atlas, we called it NEWAT (new atlas). It was already clear to us that sequence databases were not mere repositories of data; rather, they lead to new knowledge in the form of matches made.

Only slightly scarred by earlier forays, I developed a more cautious approach about likely homologies. I wrote a well-received article in *Science* entitled 'Similar amino acid sequences: chance or common ancestry?' The article (Doolittle, 1981) contained a long list of known and alleged homologies, assessed by the device of gauging resemblances by also comparing scrambled sequences, as had been shown to be effective by workers at the NBRF. The analysis was made possible on two counts: (1) some limited access to the Chemistry Department's VAX computer, and (2) the use of a great shortcut in the Needleman–Wunsch approach programmed by my older son. The article also suggested that it should be possible to trace the ancestry of proteins back to a small starter set of sequences.

In 1981 we obtained a link to the Chemistry Department's VAX computer. Armed with this magnificent new tool and an ever-growing sequence collection, we intensified the searching campaign. It was exciting. Every new entry was a chance to learn some new connection. I found myself constantly scrutinizing the output like some financier who can't tear himself away from the ticker tape, watching for the telltale asterisks with which we flagged likely matches. Some of these matches proved to be extremely important, including several that correlated oncogenes with normal cellular components, and many

others that provided wholly unexpected evolutionary connections (Doolittle, 1997).

Word quickly spread of our willingness to help anyone who contacted us, with no obligations or cost (except that we asked that sequences be left in the database, in a confidential mode if necessary). During the period 1982–1988 we searched several hundred sequences for other researchers. We made many of these people very happy, and I take great pride in the scores of articles which were published during this period in which personal acknowledgement is made to our help in identifying relationships.

It was not always happy, however. In 1984, for example, a very awkward situation arose when *Nature* sent me a manuscript to review. The authors had been sequencing some peptides from the blood coagulation proteins factors V and VIII, which were already thought to be homologous on other grounds. In their article the authors now claimed that the sequences were not only homologous to each other, but also to the α chain of fibrinogen, something I knew could not be true (at the time, our laboratory having just completed the 610-residue sequence, I had virtually memorized it). To make the point, I searched their sequences against our NEWAT database, expecting to show that any match with the fibrinogen chain would not exceed background. To my astonishment, I found that the sequences were clearly homologous to the copper-binding protein ceruloplasmin! What to do about it? I told *Nature* exactly what I did and suggested the authors add an addendum to their paper. I read over their policy about confidentiality of reviewers and told them that, should the authors have any questions, *Nature* could disclose my identity or not as they saw fit.

A few weeks later I was stunned when I signed for a certified letter from the two principal authors of the paper, who were accusing me of 'an [appalling] breach of ethics' and ordering me 'to remove (their) data or congenors thereof from any data storage device and not to divulge the data or relationships ... etc., etc ... to any person or company'. I was flabbergasted. Apparently, what had happened was that *Nature*, without my knowledge, had written to the authors and suggested that I be included as a co-author on a revised report! This was a terrible situation. The resolution was extremely unsavory, and when I retire some day, I plan to write more about this and other skulduggery in a more detailed set of memoirs. John Maddox, editor of *Nature*, was so shaken by the incident that he mentioned it in two separate News and Views later that year (Maddox, 1984a,b). In the first, he wrote that '*Nature* was still bruised by the angry withdrawal of an important article through the innocent transmittal of a referee whose interpretation of the data was more interesting than the authors'. In the second he went into more detail (Maddox, 1984b).

Multiple alignments

Beginning in 1983, we put a great deal of effort into developing a simple system for constructing phylogenetic trees from protein sequences. The 'we' in this case included my now long-time research associate, Da-Fei Feng, and Mark Johnson, who was a graduate student with me from 1983 to 1987. When we began the project it was possible to make a phylogenetic tree by the method of Fitch and Margoliash (1967) if one (a) already had a decent multiple alignment, (b) a program for finding the branching order, and (c) another for determining the branch lengths. There was also the problem of what to do with the negative branch lengths that frequently emerged when one tried to accommodate all the data in a pairwise distance matrix. Our aim was to combine all operations into a single routine that could be called up simply by typing 'tree'. It was a long and rocky road, and many others found smoother highways. But we eventually achieved our goal (Feng and Doolittle, 1996).

It slowly dawned on us that we weren't necessarily interested in the mathematically optimal alignment of a set of sequences. In fact, Mark Johnson had written a program that could optimally align up to five sequences (Johnson and Doolittle, 1986). The problem was, Mark found, that given a set of homologous sequences, the gaps showed up at different places depending on whether the sequences were examined in subsets of two, three or four. Obviously, they couldn't all be correct. We realized that what we needed was a historically accurate alignment. As such, we felt we should align the two most similar sequences first. Then, before adding the next sequence, we should freeze any gaps in the first pair. After that, the next most similar sequence was compared with an average of the first two, and so forth. The thrust of this progressive alignment scheme was embodied in the phrase: 'once a gap, always a gap' (Feng and Doolittle, 1987). The method was not only biologically sensible, it was computationally much easier than trying to effect a global alignment.

A number of other multiple alignment schemes made their appearance at about the same time (Barton and Sternberg, 1987; Taylor, 1987, *inter alia*). The progressive approach was also speeded up and improved by some simple but effective modifications that allowed multiple alignments even on microcomputers (Corpet, 1988; Higgins and Sharp, 1988). As is wont to happen in our competitive world, there was some grumbling about priority claims. Some of the grumbles suggested that the idea had been around for years, and a paper by Sankoff (1975) has been mentioned as containing the seeds of the process (Higgins *et al.*, 1996). I have since searched out that paper and gone through it carefully. If the seeds of progressive alignment are there, they are deeply buried indeed, which may account for the very long germination time.

Of authors and readers

In 1986 I wrote a thin book, the short title of which was '*URFs and ORFs.*' It was a primer in the literal sense, a low level introduction to analyzing sequences. It was a simple book intended for experimentalists like myself. Sales got a big boost when Walter Fitch wrote a review for *Cell*, which ended with the caveat '...definitely for tyros'. What Walter underestimated, I think, was how many molecular biologists qualified as *tyros* in those days.

A few years later I was asked to edit a volume of *Methods in Enzymology* on much the same subject. First I had to find and cajole 40 or so authors willing to write chapters. Then I had to read what they wrote. It was an eye-opener for me. Many of the articles were at a level well beyond either my capability or my interest. Over the years, I had tried to fathom the writings of the more theoretically inclined phylogeneticists, but their articles always left me in a dizzying maze of edges and vertices. They were clearly thinking on an elevated plane. Now I realized that I was well out of my depth in this field and should retreat to some area where I could at least understand the terminology.

But the world operates in strange ways. The volume was a great success, and I was soon beset upon to edit a follow-up. Although I hesitated, I didn't have the courage to decline. So we went through the exercise again. I can only say that if I had been able to retain everything I read in all those chapters, I would likely qualify as a bioinformaticist. But I didn't.

The human genome initiative

It was the Human Genome Initiative that gave rise to the age of bioinformatics. In the late 1980s, I was a member of one of the many committees asked to advise on whether or not this endeavor should be underaken. At the time there was a good deal of opposition to the idea. Some felt that it would compromise the small individual investigator kind of science (they were probably correct). Others worried that useful information might not emerge. In particular, there was great concern that it might not be possible to identify genes on the basis of raw genomic DNA sequence alone.

In this regard, I was surprised to find that many biomedical scientists did not appreciate the evolutionary aspect. Nor did they understand why sequence searching had already been so effective in finding relationships. That all living things were the result of a vast expansion of genes by duplication was not something they had given much thought to. As more and more gene families were uncovered, the tide gradually changed.

The human genome initiative, what with its logical spread to the study of other genomes, has affected every aspect of biology and much of medicine. The network of lineages relating all living creatures is being revealed. Nowhere has the Darwinian notion of descent with modification been more dramatically illustrated and the biochemical unity of life made clearer. The Book of Life is opening up before us for all to enjoy.

It is all being made possible by the remarkable speed and capacity of modern computers and the fantastic software that has co-evolved with them. Most impressive of all is the instant access provided to every interested investigator. Every biologist can search a new sequence the moment it emerges from the bench. There is no longer a need for a middle man. Places like the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI) devise and maintain searching programs that are wonderfully user-friendly. Not surprisingly, I am feeling a bit redundant, and I have sought refuge in the area of protein crystallography, where no one expects me to be expert. Now, when I'm not worrying about getting synchrotron time, I mostly use the databases to revisit my early project on the evolution of blood clotting proteins.

Nonetheless, I am grateful to the editors of *Bioinformatics* for inviting me to participate in this special History issue and hope readers will forgive its anecdotal style and rather personal perspective. Some individuals might take umbrage at certain of my remarks. But science is not all thrill and satisfaction. There is always disappointment and frustration, and occasionally some outright pain. Sparring often occurs even between friends.

Acknowledgements

I am grateful to all of those who helped me with my early computer studies, including especially Karen Anderson, Larry Doolittle, Will Doolittle, Da-Fei Feng, Mark Johnson and Neal Woodbury.

References

Barnett,D.R., Lee,T.-H. and Bowman,B. (1972) Amino acid sequence of the human haptoglobin β chain. Amino and carboxyl-terminal sequences. *Biochemistry*, **11**, 1189–1194.

Barton,G.J. and Sternberg,M.J. (1987) A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. *J. Mol. Biol.*, **198**, 327–387.

Blomback,B., Blomback,M., Grondahl,N.J. and Holmberg,E. (1966) Structure of fibrinopeptides—its relation to enzyme specificity and phylogeny and classification of species. *Arkiv Kemi*, **25**, 411–428.

Both,G.W. and Sleigh,M.J. (1980) Complete nucleotide sequence of the haemagglutinin gene from human influenza virus of the Hong Kong subtype. *Nucleic Acids Res.*, **8**, 2561–2575.

Brew,K., Vanaman,T.C. and Hill,R.L. (1967) Comparison of the amino acid sequence of bovine α-lactablumin and hens eggwhite lysozyme. *J. Biol. Chem.*, **242**, 3747–3749.

Chou,P.Y. and Fasman,G.D. (1974) Prediction of protein conformation. *Biochemistry*, **13**, 222–245.

Corpet,F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.*, **16**, 10 881–10 890.

Dayhoff,M.O. and Eck,R.V. (1968) *Atlas of Protein Sequence and Structure, 1967–1968*. Chapter 4, National Biomedical Research Foundation, Washington D.C..

Doolittle,R.F. (1970) Evolution of fibrinogen molecules. *Thromb. Diath. Haem.*, **Suppl. 39**, 25–42.

Doolittle,R.F. (1979) Protein evolution. In Neurath,H. and Hill,R.L. (eds), *The Proteins* Vol. IV, Academic Press, New York, pp. 1–115.

Doolittle,R.F. (1981) Similar amino acid sequences: chance or common ancestry? *Science*, **214**, 149–159.

Doolittle,R.F. (1986) *Of URFS and ORFS: A Primer on How to Analyze Derived Amino Acid Sequences*. University Science Books, Mill Valley, California.

Doolittle,R.F. (1997) Some reflections on the early days of sequence searching. *J. Mol. Med.*, **75**, 239–241.

Doolittle,R.F. and Blomback,B. (1964) Amino acid sequence investigations of fibrinopeptides from various mammals: evolutionary implications. *Nature*, **202**, 147–152.

Doolittle,R.F., Oncley,J.L. and Surgenor,D.M. (1962) Species differences in the interaction of thrombin and fibrinogen. *J. Biol. Chem.*, **237**, 3123–3127.

Doolittle,R.F., Goldbaum,D.M. and Doolittle,L.R. (1978) Designation of sequences in the 'coiled coil' interdomainal connections in fibrinogen: construction of an atomic scale model. *J. Mol. Biol.*, **120**, 311–325.

Doolittle,R.F., Watt,K.W. K., Cottrell,B.A., Strong,D. and Riley,M. (1979) Amino acid sequence of the α chain of human fibrinogen. *Nature*, **280**, 464–468.

Eck,R.V. and Dayhoff,M.O. (1965) *Atlas of Protein Sequence and Structure, 1965*. National Biomedical Research Foundation, Silver Spring, Maryland.

Feng,D.-F. and Doolittle,R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, **25**, 351–360.

Feng,D.-F. and Doolittle,R.F. (1996) Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. In Doolittle,R.F. (ed.), *Methods in Enzymology* Vol. 266, Academic Press, New York, pp. 368–382.

Fitch,W.M. (1966) An improved method for testing for evolutionary homology among proteins. *J. Mol. Biol.*, **16**, 9–16.

Fitch,W.M. and Margoliash,E. (1967) Construction of phylogenetic trees. *Science*, **155**, 279–284.

Friedmann,T., Doolittle,R.F. and Walter,G. (1978) Amino acid sequence homology between polyoma and SV40 tumor antigens deduced from nucleic acidsequences. *Nature*, **274**, 291–293.

Higgins,D.G. and Sharp,P. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–244.

Higgins,D.G., Thompson,J.D. and Gibson,T.J. (1996) Using CLUSTAL for multiple sequence alignments. In Doolittle,R.F. (ed.), *Methods in Enzymology* Vol. 266, Academic Press, New York, pp. 383–402.

Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA*, **78**, 3824–3828.

Johnson,M.S. and Doolittle,R.F. (1986) A method for the simultaneous alignment of three or more amino acid sequences. *J. Mol. Evol.*, **23**, 267–278.

Jue,R.A., Woodbury,N.W. and Doolittle,R.F. (1980) Sequence homologies among *E. coli* ribosomal proteins: evidence for evolutionarily related groupings and internal duplictions. *J. Mol. Evol.*, **15**, 129–148.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105.

Maddox,J. (1984a) Perils of too much disclosure. *Nature*, **309**, 665.

Maddox,J. (1984b) Privacy and the peer-review system. *Nature*, **312**, 497.

Mross,G.A. and Doolittle,R.F. (1967) Amino acid sequence studies on artiodactyl fibrinopeptides: II. Vicuna, elk, muntjak, pronghorn antelope and water buffalo. *Arch. Biochem. Biophys.*, **122**, 674–684.

Murata,M., Richardson,J.S. and Sussman,J.L. (1985) Simultaneous comparison of three protein sequences. *Proc. Natl. Acad. Sci. USA*, **82**, 3073–3077.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search of similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Rose,G.D. (1978) Prediction of chain turns in globular proteins on a hydrophobic basis. *Nature*, **272**, 586–590.

Sankoff,D. (1975) Minimal mutation trees of sequences. *SIAM J. Appl. Math.*, **28**, 35–42.

Taylor,W.R. (1987) Multiple sequence alignment by a pairwise algorithm. *CABIOS*, **3**, 81–87.

Walter,G., Scheidtmann,K.-H., Carbone,A., Laudano,A.P. and Doolittle,R.F. (1980) Antbodies specific for the carboxy- and amino-terminal regions of simian virus 40 large tumor antigen. *Proc. Natl. Acad. Sci. USA*, **77**, 5197–5200.

Wu,H., Lustbader,J.W., Liu,Y., Canfield,R.L. and Hendrickson,W.A. (1994) Structure of human chorionic gonadotropin at 2.6 Å resolution from MAD analysis of the selenomethionyl protein. *Structure*, **2**, 545–558.