

Introduction to sequence comparisons - why? April 3, 2002, D. Gusfield

Before we start, I want to emphasize a terminology distinction between *substrings* and *subsequences*. A subsequence differs from a substring in that the characters in a substring *must* be contiguous, while the characters in a subsequence embedded in a string need not be. For example, the string *xyz* is a subsequence, but not a substring, in *axayaz*.

1 The importance of (sub)sequence comparison in molecular biology

Sequence comparison, particularly when combined with the systematic collection, curation and search of databases containing biomolecular sequences, has become essential in modern molecular biology. Commenting on the (then) near-completion of the effort to sequence the entire yeast genome (now finished), Stephen Oliver is quoted [12]:

In a short time it will be hard to realize how we managed without the sequence data. Biology will never be the same again.

One fact explains the importance of molecular sequence data and sequence comparison in biology.

The first fact of biological sequence analysis

The *first fact of biological sequence analysis*: In biomolecular sequences (DNA, RNA or amino acid sequences), high sequence similarity usually implies significant functional or structural similarity.

The *first fact* is a by-product of *common descent*, the fact that modern (extant) molecules in different species and populations have common origins (common ancestors), and that some details of the common ancestors are preserved in extant molecules. In the jargon of evolution, two molecules are *homologous* if they descend from a common ancestor. Well, actually, if you go back far enough all of life is thought to descend from some common ancestor, so in practice, two molecules are considered homologous, if they descend from a *relatively recent* common ancestor. That's vague, but will do for now. (By the way *homology* is "the state of being homologous" in the Dictionary of Genetics.)

Homology (recent common descent) can be deduced from observations of extant organisms and molecules because evolution reuses, builds on, duplicates and modifies those structures (proteins, exons, DNA regulatory sequences, morphological features,

enzymatic pathways, etc.) that have been “successful” (left as a vague concept). Life is based on a repertoire of structured and interrelated molecular building blocks that are shared and passed around. The same and related molecular structures and mechanisms show up repeatedly in the genome of a single species, and across a very wide spectrum of divergent species. “Duplication with modification” [3, 6, 4, 5] is the central paradigm of protein evolution, wherein new proteins and/or new biological functions are fashioned from earlier ones. R. Doolittle [5, 6] emphasizes this point as follows:

The vast majority of extant proteins are the result of a continuous series of genetic duplications and subsequent modifications. As a result, redundancy is a built-in characteristic of protein sequences, and we should not be surprised that so many new sequences resemble already known sequences.

and

“... all of biology is based on an enormous redundancy ...”.

The following quotes reinforce this view and suggest the utility of the “enormous redundancy” in the practice of molecular biology. The first quote is from Eric Wieschaus, co-winner of the 1995 Nobel prize in medicine for work on the genetics of *Drosophila* development. The quote is taken from an Associated Press article October 9, 1995. Describing the work done years earlier, Wieschaus says

We didn’t know it at the time, but we found out everything in life is so similar, that the same genes that work in flies are the ones that work in humans.

And fruit flies aren’t special. The following is from a book review on DNA repair [11]:

Throughout the present work we see the insights gained through our ability to look for sequence homologies by comparison of the DNA of different species. Studies on yeast are remarkable predictors of the human system!

Here are two recent quotes that I particularly like from an article in *The Scientist*, June 1998 “Flies Invade Human Genetics” by Ricki Lewis.

Back then (the 1980’s), when a gene was discovered in *Drosophila* that had a homolog in a human, it was [like] wow! Now a researcher feels apologetic if there isn’t a homology, observes Geoffrey Duyk

sequence similarities, called homologies, between many fly and human genes are 'stunning and striking' says Thomas Kaufman. 'Whole pathways in flies are maintained in humans, using the same molecules. Who would have believed 20 years ago that that would be the case? Then, we were just working on peculiar genes in fruit flies', he added.

So "redundancy", "similarity", "homology" are central phenomena in biology. But similarity has its limits – humans and flies do differ in some respects. These differences make *conserved* similarities even more significant, which in turn makes *comparison* and *analogy* very powerful tools in biology. Lesk [9] writes:

It is characteristic of biological systems that objects that we observe to have a certain form arose by evolution from related objects with similar but not identical form. They must, therefore, be robust, in that they retain the freedom to tolerate some variation. We can take advantage of this robustness in our analysis: By identifying and comparing related objects, we can distinguish variable and conserved features, and thereby determine what is crucial to structure and function.

The important "related objects" to compare include much more than sequence data, because biological universality occurs at many levels of detail. But it is usually easier to acquire and examine sequences than it is to examine fine details of genetics or cellular biochemistry or morphology. For example, there are vastly more protein sequences known (deduced from underlying DNA sequences) than there are known three-dimensional protein structures. And it isn't just a matter of convenience that makes sequences important. Rather, the biological sequences *encode* and reflect the more complex common molecular structures and mechanisms that appear as features at the cellular or biochemical levels. Moreover, "nowhere in the biological world is the Darwinian notion of 'descent with modification' more apparent than in the sequences of genes and gene products" [6]. Hence a tractable, though partly heuristic, way to search for functional or structural universality in biological systems is to search for similarity and conservation at the *sequence* level. The power of this approach is made clear in the following quotes from [10] and [1] respectively:

Today, the most powerful method for inferring the biological function of a gene (or the protein that it encodes) is by sequence similarity searching on protein and DNA sequence databases. With the development of rapid methods for sequence comparison, both with heuristic algorithms and powerful parallel computers, discoveries based solely on sequence homology have become routine.

Determining function for a sequence is a matter of tremendous complexity, requiring biological experiments of the highest order of creativity. Nevertheless, with only DNA sequence it is possible to execute a computer-based algorithm comparing the sequence to a database of previously characterized genes. In about 50% of the cases, such a mechanical comparison will indicate a sufficient degree of similarity to suggest a putative enzymatic or structural function that might be possessed by the unknown gene.

So large-scale sequence comparison, usually organized as database search, is a very powerful tool for biological inference in modern molecular biology. And that tool is almost universally used by molecular biologists. It is now standard practice, whenever a new gene is cloned and sequenced, to translate its DNA sequence into an amino acid sequence and then search for similarities between it and members of the protein databases. No one today would even think of publishing the sequence of a newly cloned gene without doing such database searches.

The final quote reflects the potential total impact on biology of the *first fact* and its exploitation in the form of sequence database searching. It is from an article [7] by Walter Gilbert, Nobel prize winner for the co-invention of a practical DNA sequencing method.

The new paradigm now emerging, is that all the ‘genes’ will be known (in the sense of being resident in databases available electronically), and that the starting point of biological investigation will be theoretical. An individual scientist will begin with a theoretical conjecture, only then turning to experiment to follow or test that hypothesis.

Already, hundreds (if not thousands) of journal publications appear each year that report biological research where sequence comparison and/or database search is an integral part of the work. Many such examples that support and illustrate the *first fact* are distributed throughout the class. But before discussing those examples, we must first develop, the techniques used for approximate matching and (sub)sequence comparison.

Caveat

The *first fact of biological sequence analysis* is extremely powerful, and its importance will be further illustrated throughout the book. However, there is not a one-one correspondence between sequence and structure or sequence and function, because the *converse* of the *first fact* is *not* true. That is, high sequence similarity usually implies significant structural or functional similarity (the first fact), but structural or functional similarity does *not* necessarily imply sequence similarity. On the topic of protein

structure, F. Cohen [2] writes “... similar sequences yield similar structures, but quite distinct sequences can produce remarkably similar structures.” This *converse* issue is discussed in greater depth when we examine multiple sequence comparison.

1.1 The double-Muscling Story - the use of bioinformatics in a gene cloning effort

See Ref 8 by S. Dickman in the suggested paper list linked from the class homepage.

2 Much of current bioinformatics concerns sequence comparisons and alignments

So what we want sequence comparison methods to do, is to reveal evidence of homology, and do it in a way that exposes the important shared and diverged features. Many biological inferences can then be made from what is shared (conserved) and what is diverged. When we look at specific definitions for alignment, and specific method for computing alignments, we should always keep in mind what alignments are for and what we want to learn from them; then ask whether the alignment objective functions and the computational methods really deliver what we want from them.

The area of approximate matching, and sequence comparison is central in bioinformatics/computational molecular biology both because of the presence of errors in molecular data, and because of active mutational processes that (sub)sequence comparison methods seek to model and reveal. On the *technical* side, sequence *alignment* has become the central tool for sequence *comparison* in molecular biology. Henikoff and Henikoff [8] write:

Among the most useful computer-based tools in modern biology are those that involve sequence alignments of proteins, since these alignments often provide important insights into gene and protein function. There are several different types of alignments: global alignments of pairs of proteins related by common ancestry throughout their lengths, local alignments involving related segments of proteins, multiple alignments of members of protein families, and alignments made during data base searches to detect homologies.

This statement provides a framework for much of our examination of sequence alignment. We will examine in detail the four types of alignments (and several variants) mentioned above. And we will show how those different alignment models address different kinds of problems in biology.

References

- [1] C. Caskey, R. Eisenberg, E. Lander, and J. Straus. Hugo statement on patenting of dna. *Genome Digest*, 2:6–9, 1995.
- [2] F. E. Cohen. Folding the sheets: Using computational methods to predict the structure of proteins. In E. Lander and M. S. Waterman, editors, *Calculating the Secrets of Life*, pages 236–271. National Academy Press, 1995.
- [3] R. F. Doolittle. *Of Urfs and Orfs: A primer on how to analyze derived amino acid sequences*. University Science Books, Mill Valley, CA., 1986.
- [4] R. F. Doolittle. Redundancies in protein sequences. In G. Fasman, editor, *Prediction of protein structure and the principles of protein conformation*, pages 599–624. Plenum, 1989.
- [5] R. F. Doolittle. Searching through sequence databases. In R. F. Doolittle, editor, *Methods in Enzymology vol. 183. Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, pages 99–110. Academic Press, 1990.
- [6] R. F. Doolittle. What we have learned and will learn from sequence databases. In G. Bell and T. Marr, editors, *Computers and DNA*, pages 21–31. Addison-Wesley, 1990.
- [7] W. Gilbert. Towards a paradigm shift in biology. *Nature*, 349:99–99, 1991.
- [8] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. of the Nat. Academy of Science*, 89:10915–10919, 1992.
- [9] A.M. Lesk. Computational molecular biology. In A. Kent and J.G. Williams, editors, *Encyclopedia of Computer Science and Technology, Vol. 31*, pages 101–165. Marcel Dekker, 1994.
- [10] William R. Pearson. Protein sequence comparison and protein evolution. Tutorial T6 of Intelligent Systems in Mol. Bio., Cambridge England, July 1995.
- [11] B. S. Strauss. Book review: Dna repair and mutagenesis. *Science*, 270:1511–1513, 1995.
- [12] N. Williams. Closing in on the complete yeast genome sequence. *Science*, 268:1560–1561, June 1995.