Mapping and Sequencing the Human Genome:

A Beginner's Guide to the Computational Science Perspective

by *Wayne Smith*

Introduction: Data, Data Everywhere and Not a Drop to Drink

Before the 1950's, geneticists studied the mechanisms of how genes are passed from one generation to the next by using cross-breeding experiments. However, the discovery of the molecular structure of the genetic material and its chemical properties opened up a fertile field for understanding biology through the fields of chemistry and physics, without using live organisms. As scientists probed into the chemical makeup of genes, their growing understanding helped them develop well-defined processes for extracting genetic information. This produced volumes of data growing far more quickly than could be analyzed. Computational science has come alongside these efforts to decipher the genetic machinery of humans.

The next section provides an overview of what is involved in deciphering a genome. After defining the problem, this paper surveys significant areas where computational science has helped to solve the problem. Though scientists regularly use computers to assist in laboratory methods for data collection and verification, the main developments from the computational science perspective are in data organization, access and analysis. The last section concludes by discussing emerging areas for future research.

Cracking the Genetic Code: The Problem Defined

Mapping and sequencing a genome consists of discovering how biological traits are encoded in the genetic material of an organism. Many biological characteristics can be traced to one or more genes. A **gene** is a portion of a **DNA (Deoxyribonucleic Acid)** strand. This strand is made up of millions (sometimes billions) of smaller molecules called **nucleotides**. **Mapping** gives an estimate of the location of a gene in the DNA. **Sequencing** identifies the exact nucleotides that make up a gene.

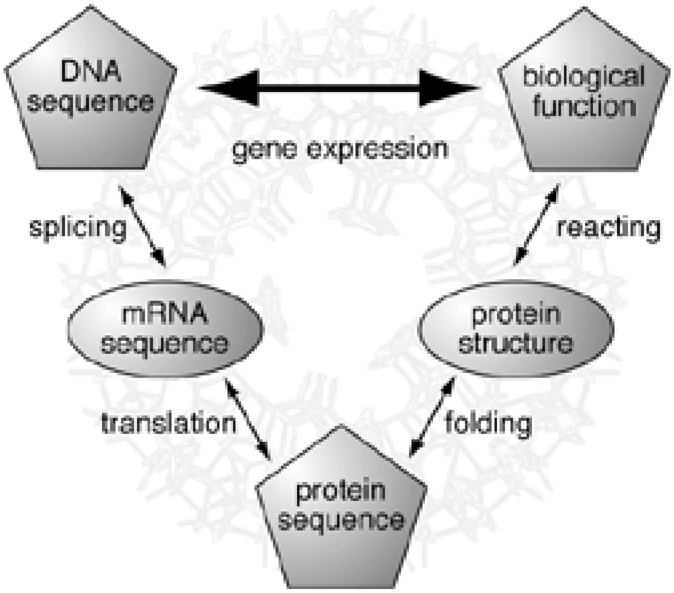
Organisms are made out of cells, and cell behavior is governed by its genetic makeup. Various chemicals act as agents to activate genes in a process called **gene expression**. Once a gene has been ``turned on'', it undergoes transcription and translation. **Transcription** uses the DNA as a template to create a mRNA molecule, which is then **translated** into proteins made up of amino acids. As proteins are formed from the DNA, they spontaneously fold into three dimensional shapes that dictate their chemical function. This process, from encoded gene to expressed gene, frames the problem of deciphering the genome. Scientists want to associate a particular portion of DNA with a particular biological function. By decoding the DNA, they hope to define a unified theory of biology that explains all biological functions.

Mapping and Sequencing Sub-problems

As scientists break down the gene expression process, several key sub-problems have emerged, as illustrated in Figure 1. While trying to find direct associations between DNA and biological functions, several questions arose:

- How do common DNA sequences relate to the results of transcription?
- What is the correlation between transcription and translation?
- What are the associations between translation results and protein folding?

DNA and amino acid arrangements represent *sequence* data, and protein structures (called **conformations**) represent the *structure* data. Functional data that relates the two exist in medical and pharmaceutical databases, but functional analysis is still relatively new (see this paper's conclusion). The whole problem can be summarized as trying to relate patterns among the sequence data to patterns among the structures that, in turn, map to patterns among the protein functions. Scientists are hunting for the basic principles and mechanisms that effectively explain the process of gene expression. The hunt for principles has ranged from *knowledge-intensive* methods that use chemistry and quantum mechanics, to *data-intensive* methods that perform statistical analysis of existing sequence and structure data.



Finding a Nucleotide in a Haystack: Data Organization and Access

As scientists continue to collect data, a multitude of databases have emerged ([7, 10] survey many current databases). Some databases hold information on a particular organism while others specialize in a specific type of information (e.g. only sequences or only bibliographies). One of the primary sequence databases is the *Genome Database* at Johns Hopkins (found at http://gdbwww.gdb.org, and it is replicated at nine international sites. Two important structure databases are Brookhaven's *Protein Data Bank (PDB)* (found at http://pdb.pdb.bnl.gov and Cambridge's *Crystallographic Data Bank* (found at http://csdvx2.ccdc.cam.ac.uk. Apart from simply collecting the information on-line, efforts to improve infrastructure have focused on finding common schemas for the many databases and on making databases more accessible (e.g. via the web) and more usable. It is critical that the genetic data be intuitively presented to its users: geneticists, molecular biologist, etc.

Promising Patterns: Sequence Analysis

Between transcription and translation, portions of the DNA are spliced out and never translated into protein. Some of the earliest programs, therefore, searched DNA sequences for areas that would not be spliced out. Statistical models of **coding** (areas not spliced out) versus **non-coding** regions have attained 95% accuracy while models using neural networks have attained 99% accuracy [3]. The current emphasis for site analysis is on finding boundaries between coding and non-coding regions.

Once the sequence databases grew to a significant size, programs for **sequence matching** emerged in the 1980's to find similarities among the collection. Today the main programs used for sequence matching are BLAST (Basic Local Alignment Search Tool) (found at http://www.ncbi.nlm.nih.gov/BLAST and FASTA (see, for example http://www.genome.ad.jp/SIT/FASTA.html). Each uses a scoring mechanism to track matches of subsequences, and the intermediate scores narrow the search for larger subsequence matching [14]. These techniques and their variations are capable of classifying approximately 40% to 50% of the new sequences. The remainder falls into what is commonly called the *Twilight Zone* of sequence similarities because their relationships are too subtle for pure sequence comparison methods to determine. To make further classifications, scientists use structure data or rely on principles from chemistry and physics.

Welcome to the Fold: Structure Analysis

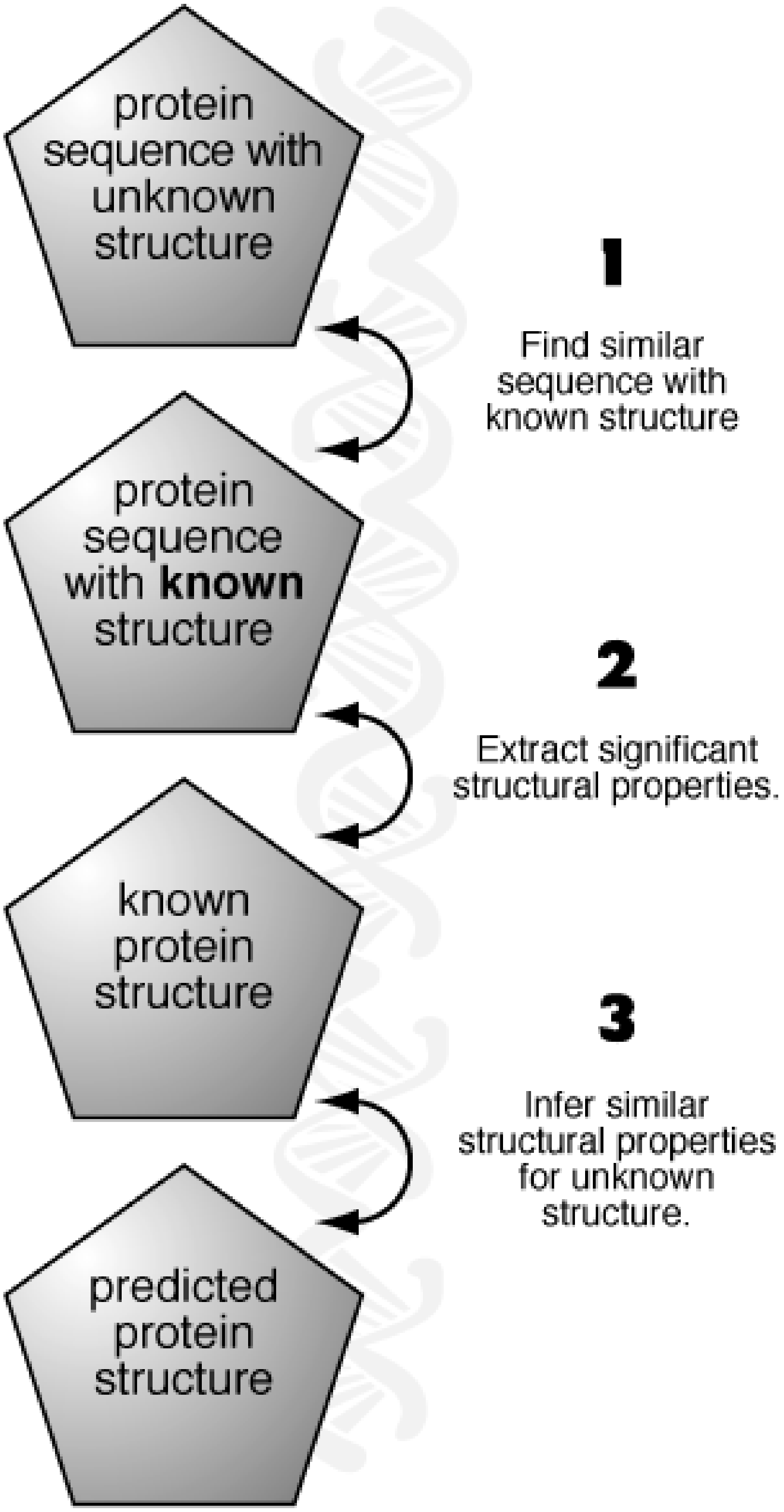
Quantum mechanics provides a foundation for deriving the shape of a molecule based on its constituent atoms. However, the calculations required for a molecule of even only six atoms are enormous and requires many assumptions. Nevertheless, simulations of molecular dynamics have been used for years to understand protein structures [4]. Scientists have also formulated ways to compare and group protein structures using the two- and three-dimensional coordinates of the constituent atoms. These and other techniques are described in more detail in [2].

Those Confounding Conformations: Structure Prediction

A protein has a primary, secondary, tertiary and quaternary structure. Some recent analyses use six levels that add super-secondary structure (between secondary and tertiary) and domains (between super-secondary and tertiary).

- The primary structure of a protein is the most basic description consisting of its sequence of amino acids.
- The secondary structure accounts for identifiable folding patterns emerging from small segments of the protein strand. Secondary structures have traditionally been categorized into three classes: alpha helices (see http://www.pdb.bnl.gov/PPS2/course/section8/ss-960531_6.html) , beta sheets (see, http://www.pdb.bnl.gov/PPS/course/3_geometry/sheet.html), and random coils (see http://www.pdb.bnl.gov/PPS2/course/section8/ss-960531_17.html).
- The tertiary structure is the final conformation that results when the newly formed protein sequence folds into its lowest energy state (the state that is most stable). In some cases, multiple tertiary structures may combine in a quaternary structure to form a functional protein. For example, hemoglobin is an important protein in our blood. In human hemoglobin, several hundred amino acids (primary) arrange themselves in a series of alpha helices, beta sheets and random coils (secondary) to form a tertiary unit sub-unit. Four of these sub-units combine (together with an iron atom) to form the hemoglobin in our blood (quaternary).

Around 1961 researchers began to understand that protein sequence largely determines protein structure. Scientists have tried a multitude of techniques to associate the primary structure with the secondary and tertiary structures. Early attempts from the 1960's, 1970's and 1980's used statistical correlation, dynamic programming and other techniques, all without significant progress. Meanwhile, in the laboratories, the determination of protein structures continues to rely on expensive and time-consuming processes such as X-ray crystallography. There are many undiscovered sequences, but there are far more unknown structures. Sequence comparison and classification allow a method for structure prediction. Given a known sequence with an unknown structure, one may find its structure using known structures of similar sequences (see Figure 2). This basic approach is used in almost all of the recent predictive methods.


protein sequence with unknown structure
1
Find similar sequence with known structure
protein sequence with **known** structure
2
Extract significant structural properties.
known protein structure
3
Infer similar structural properties for unknown structure.
predicted protein structure

Patterns that Predict

Scientists have pursued sequence-to-structure prediction using molecular dynamics simulations, simulated annealing, Markov Models, stochastic grammars, neural networks and heuristic searches.

- **Molecular dynamics simulations** use energy minimization calculations to determine whether a known sequence will take the shape of a known structure [8].

- Another approach uses **simulated annealing** of folding potentials to derive conformation from a sequence. Simulated annealing begins with a primary sequence and a simulated starting temperature [11]. It iteratively steps the temperature down on a specified cooling schedule and determines if a global energy minimum emerges among a collection of possible conformations. Conformations are eliminated each iteration using chemical attributes relevant to a global minimum energy state.
- **Neural networks** accomplish structure prediction by initially training a network with known sequence-structure associations [12]. The network then takes a known sequence and predicts its structure based on its training.
- General artificial intelligence search techniques (such as in [5, 6]) predict protein structure by capturing and using complex patterns representing essential sequence-to-structure associations for a category of proteins. The traditional matching algorithm of **Hidden Markov Models (HMM)**, called the Baum-Welch Algorithm (also called the Forward-Backward algorithm), uses probabilities to correlate amino acid positions. These correlations enable predictions that certain amino acids will appear as neighbors in a sequence.
- **Stochastic linear grammars** (or regular grammars) are exactly equivalent to HMMs. Stochastic context free grammars claim an improvement over HMMs because HMMs do not take into account pair-wise interaction of nucleotides [9]. HMMs assume fewer constraints on nucleotide distribution. One stochastic context free grammar, **Ranked Node Rewriting Grammars (RNRG)** shows some promise [1]. RNRGs came from **Tree Adjoining Grammars** (developed around 1985) but allow replacement of portions of the derivation tree based on statistical weights. This method fared well with beta-sheets that were <25% similar in primary sequence (in the *Twilight Zone*). Beta-sheets are harder to find than alpha-helices because they typically range over discontinuous sections of the sequence.
- One system combines a neural network with a mini-knowledge base containing structure predicting rules and a statistical prediction method [15]. This hybrid system makes a final prediction based on a combination of methods. One study found **Bayesian networks**, which use Bayes' rule of statistics to model uncertainty in reasoning, equally useful in predicting protein structure as neural networks [13]. However, even the best techniques at this time are getting around 60% to 65% accuracy. The early techniques from 1974 to 1988 averaged between 50% to 59%. When neural networks began to be used in 1988, the best technique attained 64%. The hybrid system of [15] claims 66.4% accuracy.

Part of the difficulty in predicting structure from sequence is that the data is sparse and unevenly distributed across all possible sequences and structures. For example, there are an estimated 90,000 protein structures associated with the human genome, but there are only about 3,000 captured in databases [2]. Predictive algorithms are constrained by the sufficiency of past associations. Sparse history leads to weak predictive power.

Conclusion

We have toured the landscape of mapping and sequencing the human genome. The basic problem is summarized as building associations between genes encoded in long DNA sequences and biological functions dictated by activating those genes. Geneticists have developed numerous techniques to collect, validate and distribute data for DNA sequences, protein sequences and protein structures. Computational scientists have developed tools for enhancing these tasks.

Working with geneticists, computational scientists will enable widely accessible databases of genetic data. Standards are being developed for common schemas, but tools will be needed to enforce those schemas and to migrate existing data to the new formats. **Data warehousing** technology for database migration and transformation should be considered for these tasks.

As the sequence and structure databases become better populated, classification and prediction algorithms will improve. The hybrid approaches appear to be the most promising. In the near- to mid-term, though, computational scientists should consider alternative approaches for predicting protein structure. The governing rule that sequence determines structure may be better applied when the sequence data is more complete. Moreover, data from protein functions and metabolic pathways should be probed, compared, classified and used to aid in structure determination. There is a wealth of information on protein and synthetic drug functions hidden in medical and pharmaceutical databases that must be tapped to build structure-to-function associations. Work is in progress, but structure-to-function associations could become the next big hurdle in deciphering genomes.

References

1

Abe, N. and Mamitsuka, H. A new method for predicting protein secondary structures based on stochastic tree grammars. In *Machine Learning: Proceedings of the Eleventh International Conference*, July 1994, Morgan Kaufmann, San Francisco, pp. 3-11.

2

Holm, L. and Sander, C. Mapping the Protein Universe. *Science* 273, 5275 (2 Aug. 1996), pp. 595-602.

3

The Human Genome Project: Deciphering the Blueprint of Heredity. Necia Grant Cooper (editor), University Science Books, Mill Valley, CA, 1994.

4

Karplus, M. and Pestko, G. Molecular dynamics simulations in biology. *Nature* 347 (18 Oct. 1990), pp. 631-639.

5

King, R. PROMIS: experiments in machine learning and protein folding. In *Machine Intelligence 12: Towards an Automated Logic of Human Thought.* Clarendon, NY, 1991, pp. 291-310.

6

Lathrop, R., Smith, T., Webster, T. and Winston, P. ARIEL: a massively parallel symbolic learning assistant for protein structure and function. In *Artificial Intelligence at MIT expanding* *frontiers*. MIT Press, 1990, pp. 70-103.

7

Organization for Economic Co-operation and Development. *The Global Human Genome Programme.* OECD Publications Service, Paris, 1995.

8

Rey, A. and Skolnick, J. Efficient algorithm for the reconstruction of a protein backbone from the alpha-carbon coordinates. *Journal of Computational Chemistry* 13, 4 (May 1992), pp. 443-456.

9

Sakakibara, Y., Brown, M., Underwood, R., Mian, I. and Haussler, D. *Stochastic Context-Free Grammars for Modeling RNA*. UCSC-CRL-93-16, Univ. of California, Santa Cruz (June 8, 1993).

10

Sillince, J. and Sillince, M. *Molecular Databases for Protein Sequence and Structure Studies*. Springer-Verlag, Berlin, 1991.

11

Snow, M. Powerful simulated-annealing algorithm locates global minimum of protein-folding potentials from multiple starting conformations. *Journal of Computational Chemistry* 13, 5 (June 1992), pp. 579-584.

12

Steeg, E. Neural networks, adaptive optimization, and RNA secondary structure prediction. In *Artificial Intelligence and Molecular Biology*, AAAI Press, Menlo Park, CA, 1993, pp. 121-160.

13

Stolorz, P., Lapedes, A. and Xia, Y. Predicting protein secondary structure using neural net and statistical methods. *Journal of Molecular Biology* 225, 1992, pp. 363-377.

14

Taubes, G. Software matchmakers help make sense of sequences. *Science* 273, 5275 (2 Aug. 1996), pp. 588-590.

15

Zhang, X., Mesirov, J. and Waltz, D. A hybrid system for protein secondary structure prediction. *Journal of Molecular Biology* 225, 1992, pp. 1049-1063.

WAYNE SMITH is a Ph.D. candidate in Computer Science at the University of South Carolina. After receiving the B.S. and M.S. in Computer Science at Clemson University, he worked with NCR Corporation in areas such as expert systems, distributed programming tools, system administration and kernel drivers and file systems. His current interests are in computational science, machine learning and intelligent agents.