

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

## **Preface: Purpose, Approach and Methods**

This report details the results of the Biomedical Computing Requirements for High Productivity Computing Systems project for DARPA's High Productivity Computing Systems (HPCS) R&D program. The biomedical HPCS community is a key segment of the industrial user community, important because of its potential market size and because of the large public-health benefits that can result from advances in medical research enabled by high-end computing. The research goal of this project is to determine biomedical computing requirements for high productivity computing systems in order to (1) define the size and nature of demand, (2) provide an assessment of the impact high productivity computing systems (HPCS) technologies can have on important biomedical problems and (3) highlight HPCS R&D areas critical to advances in biomedical computing.

The study team used multiple techniques to gather, assimilate, and validate the biomedical HPCS requirements, including: (1) a comprehensive review of biomedical computing needs as reported in current literature; (2) telephone interviews with researchers, program managers, and industry representatives; (3) a workshop to identify software environment requirements; (4) assimilation of the findings from tasks (1) - (3) into a preliminary report; (5) a final report that will synthesize the discussions from the workshop and other community input; and (6) an update to the report approximately one year later. This draft represents the results of steps (1) and (2).

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

## **1. Biomedical Computing Overview**

Biomedical computing is the application and development of computer methods for biomedical research<sup>1</sup>. It spans many disciplines including bioinformatics, molecular modeling, systems biology, medical imaging. The ultimate goal of biomedical computing is to advance the biomedical sciences by simulating life at all applicable levels of detail—the biochemical, physiological, cell, organ, organism, and population levels. The results promise to include better diagnoses, better drugs and other therapies that are developed faster, perhaps through mass customization, better surgical procedures, better prostheses, better recognition and repair of public health problems, and, thus, a healthier population, perhaps with lower medical costs.

**Example.** In collaboration with the Cleveland Clinic Foundation and the University of Auckland in New Zealand, the researchers in the Cardiac Mechanics Research Group explore the potential of a revolutionary surgical method for patients with severe heart failure. Through the combination of computational modeling with magnetic resonance imaging, the research is to predict which patients effectively can be rescued, using surgical ventricular reduction.

As has happened previously in the physical sciences, the role of computing is dramatically increasing in all areas of biological research. Although computational biology is still a relatively small subfield of biology, bioscience is widely noted as the next killer applications for high-end computational science. There is a huge variety of computational biology applications, including databases, sequence annotation, protein structure prediction, biochemical simulations, metabolic network modeling, imaging, and many others. The initial wave of computational use focused on sequence analysis. While many unsolved problems remain in sequence analysis, current and future needs will focus on integration of diverse sets of data, originating from a variety of experimental techniques which are capable of producing data at the levels of entire cells, organs, organisms and populations<sup>2</sup>.

---

<sup>1</sup> Report of the Bioinformatics Working Group of the National Advisory Research Resources Council, June 2000.

<sup>2</sup> Baldi, Pierre and Brucak, Soren. *Bioinformatics The Machine Learning Approach* MIT Press 2001. p. xi

**Federation of American Scientists**  
**Biomedical Computing Requirements for HPCS**  
**Draft Preliminary Report**

Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)

**Example.** A computational model of the cardiovascular system is aiding researchers in understanding the fundamental biochemical, biophysical, electrical and mechanical functions of the normal heart. The model is also advancing understanding of the molecular and genetic origins of heart disease, the electrical and mechanical properties of blood flow in large and small blood vessels; and the development of potential approaches for new cardiovascular drugs. A virtual lung model, developed at the Department of Energy's Pacific Northwest National Laboratory, may help predict the impact of pollutants on respiratory systems and provide new insights into asthma, as well as other pulmonary diseases<sup>3</sup>. Using the virtual respiratory tract, PNNL scientists can analyze the influence of various factors, such as the amount of pollutants or length of exposure, on healthy versus diseased lungs by manipulating the computer model. With the model they can begin to simulate how gases, vapors and particulates may act differently within lungs of people suffering from cystic fibrosis, emphysema and asthma.

Biomedical computing presents many challenges. First, biology is inherently non-linear and complex – current models are simplified linear approximations, often developed to fit into available computing resources. Inaccuracies entailed by this linearity severely limit the models' applicability.

Another challenge is the wide range of dimensions of biological interest – ranging small organic molecules to multi-protein complexes to cellular process to tissues to the interaction of human populations with the environment. For example, the alignment of two sequences of length 100 has on the order of  $10^{30}$  possible solutions. Various search strategies are employed to narrow or jump around the space, but the problem is still very hard to compute.

**Example.** Example: (Duan and Kollman 1998]. 1  $\mu$ s simulation of a 36-residue peptide starting from a fully extended state. This peptide is one of the smallest proteins that can fold autonomously, with folding estimated to take between 10  $\mu$ s and 100  $\mu$ s. It contains three short alpha-helices. The simulation involved in addition to the protein about 3000 water molecules and was performed in a truncated octahedron simulation box with a time step of 2 fs. About 4 months of computing time on a 256-processor parallel computer was required for the 1  $\mu$ s simulation. While the protein did not actually fold into the known experimental structure, a marginally stable state which showed significant resemblance to the native conformation was observed. This state had a lifetime of about 150ns.

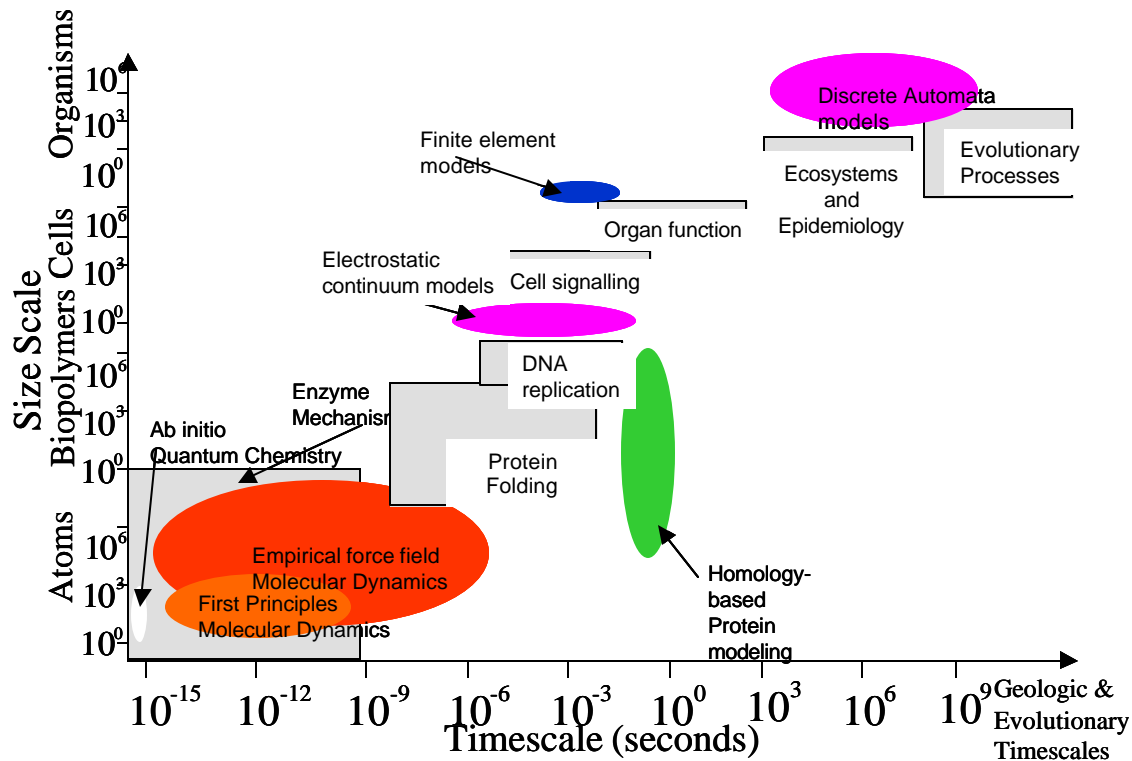
---

<sup>3</sup> <http://www.pnl.gov/news/2001/01-33.htm>

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)

A third major challenge is system complexity and the need to span multiple scales of biological organization. Timescales present yet another challenge – the timescales of biological function range from very fast femtosecond molecular motions, to multi second protein folding pathways, to cell cycle and development processes that take place over the order of minutes, hours and days, even generations. The following chart demonstrates the relationship between complexity and timescales.



**Figure 1-1 Complexity and Timescale**

There is a critical need for theoretical, algorithmic and software advances in storing, retrieving, networking, processing, analyzing, navigating and visualizing biological information. Indeed, biomedical computing is in many ways in its infancy in regards to use of computing; this presents its own set of challenges in that the field lacks well established algorithms and computational methods. In addition, while computing capabilities increase continuously,

**Federation of American Scientists**  
**Biomedical Computing Requirements for HPCS**  
**Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

biomedical computing models generally are not re-designed to take advantage of either new hardware or of the most advanced high-end systems.

The inherent complexity of biological systems, resulting from biological evolution and our lack of a comprehensive theory of biological organization at the molecular level requires sophisticated machine learning approaches in order to deal with huge amounts of data. Machine learning methods (neural networks, hidden Markov models, etc.) are well suited for domains characterized by large quantities of data, noisy patterns and the absence of general theories. These methods are computationally intensive, clearly require high-end computing capabilities, and would benefit from further improvements in computational performance.

Through interviews to date, the study team has identified the following biological research where the requirement for HPCS is already widely recognized:

- Structure of proteosome, ribozyme, ribosome, ATPases, viruses, membrane protein complexes
- Whole genome comparison
- Combined quantum/classical simulations
- Protein folding/threading
- Microsecond time-scale simulations
- Protein-protein and protein-DNA recognition and assembly

The enormous complexity of biological systems and the difficulty of using information from small model systems to address complex, collective phenomena at large scales requires significant advances in theories, algorithms, software and hardware. Currently systems allow simulations of biomolecular systems to be routinely performed for about 100 thousand atoms for tens of nanoseconds. In order to support microsecond simulations of systems with several million atoms

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

will require an increase in power by at least three orders of magnitude. The following table provides examples of current capability compared to actual problem size<sup>4</sup>.

Activity	Current Limit	Problem Size	Complexity	Memory
Ab initio study of enzyme catalysis	60 heavy atoms	250 heavy atoms	$O(n^3)$	$O(n)$
X-ray refinement of large assemblies	25,000 atoms	125,000 atoms	$O(n^2)$	$O(n^2)$
Large scale protein motion, membrane transport	200 residues	1000 residues	$O(n \log n)$	$O(n)$
Flexible docking of chemical databases	3000 compounds	1,000,000 compounds	$O(n)$	$O(n)$
Phylogenetic mapping	150 sequences	200 sequences	$O(n^3)$	$O(n \times m)$
	10,000 bases	1,000,000 bases		
RNA 3-D conformations	100 residues	1000's residues	$O(n \log n)$	$O(n)$

### **1.1. Outline of the Remainder of this Report**

This remainder of the report organizes biomedical computing in four major categories, each with a review of its algorithms and computational methods. The categories are as follows:

---

<sup>4</sup> Stan Burke. Burke, NIH presented at DARPA Biomedical Computing Requirements workshop January 17, 2003.

**Federation of American Scientists**  
**Biomedical Computing Requirements for HPCS**  
**Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

- **Bio Information Processing** (includes genomics, DNA sequencing, microarray technologies and bioinformatics): Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.
- **Computational Biology** (includes molecular modeling, tissue engineering, organ modeling and systems biology): The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological systems.
- **Protein Biochemistry** (includes protein structure and proteomics): The identification, characterization and quantification of all proteins involved in a particular pathway, organelle, cell, tissue, organ or organism that can be studied in concert to provide accurate and comprehensive data about that system.
- **Drug Discovery:** Computational approaches to lead identification and design.
- **Computer-aided diagnostic imaging and image-guided interventions :** Molecular, cellular and organ-level imaging, visualization and image analysis.
- **Integrative Modeling:** Coupling of models- physical organization, chemical, mechanical, electrical, metabolic, and thermal models. Integrative models, such as organ modeling, attempt to take into account explicit geometry of the organ, coupled to hydrodynamics, continuum mechanics, reaction-diffusion, radiation and discrete particle transport.

Some examples from Section 1 reappear where they fit in the discussions later.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

## **2. Bio Information Processing**

Molecular Bioinformatics is the development and application of computer methods for management, analysis, interpretation and prediction for molecular biology. For this report, we include genomics IT, DNA sequencing, and microarray technologies in this category. Genomics IT is complex text searching of DNA sequences used in DNA sequence assembly and analysis. There are two tasks in computational genomics: sequencing and analysis. Sequencing requires putting together millions of pieces of short error-prone sequences. Analysis of DNA sequences requires finding the individual genes within extensive non-coding regions and irrelevant base-pair sequences and other biological features (there are approximately 30,000 human genes, which comprise only 2% of the genome). Computer solutions to these problems include:

- alignment algorithms
- probabilistic techniques for sequence analysis
- large systems built from these basic algorithms

The first step in the biological hierarchy is a comprehensive genome-based analysis of the rapidly emerging genomic data. Use of new microarray-based technologies make possible high-throughput approaches that are rapidly generating terabytes of information, potentially providing fundamental insights into biological processes ranging from gene function to development, cancer, aging and pharmacology. Modern sequencers produce 1000 base pair reads/sec and operate full-time for days at a time; continual improvements in technology increase the throughput. Even partial understanding of the data can provide valuable research information. At the same time the huge quantities of data are overwhelming conventional methods of biological analysis.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

## **2.1. The Data Challenge**

Biological data is now estimated to be doubling every six months. The GenBank Database alone grew from 680,338 base pairs in 1982 to 22 billion base pairs in 2002 (compared to 13.5 base pairs as of August 2001<sup>5</sup>.) and is now doubling in around 10 months. Currently the database grows by more than 11,000,000 bases per day. (See Figure 2-1.) GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences, and is part of the International Nucleotide Sequence Database Collaboration, which is comprised of the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at the National Center for Biotechnology Information. These three organizations exchange data daily. Each GenBank entry includes a concise description of the sequence, the scientific name and taxonomy of the source organism, and a table of features that identifies coding regions and other sites of biological significance, such as transcription units, sites of mutations or modifications, and repeats. Protein translations for coding regions are included in the feature table. Bibliographic references are included along with a link to the Medline unique identifier for all published sequences.

---

<sup>5</sup> <http://www.ncbi.nlm.nih.gov/>

Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report

Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)

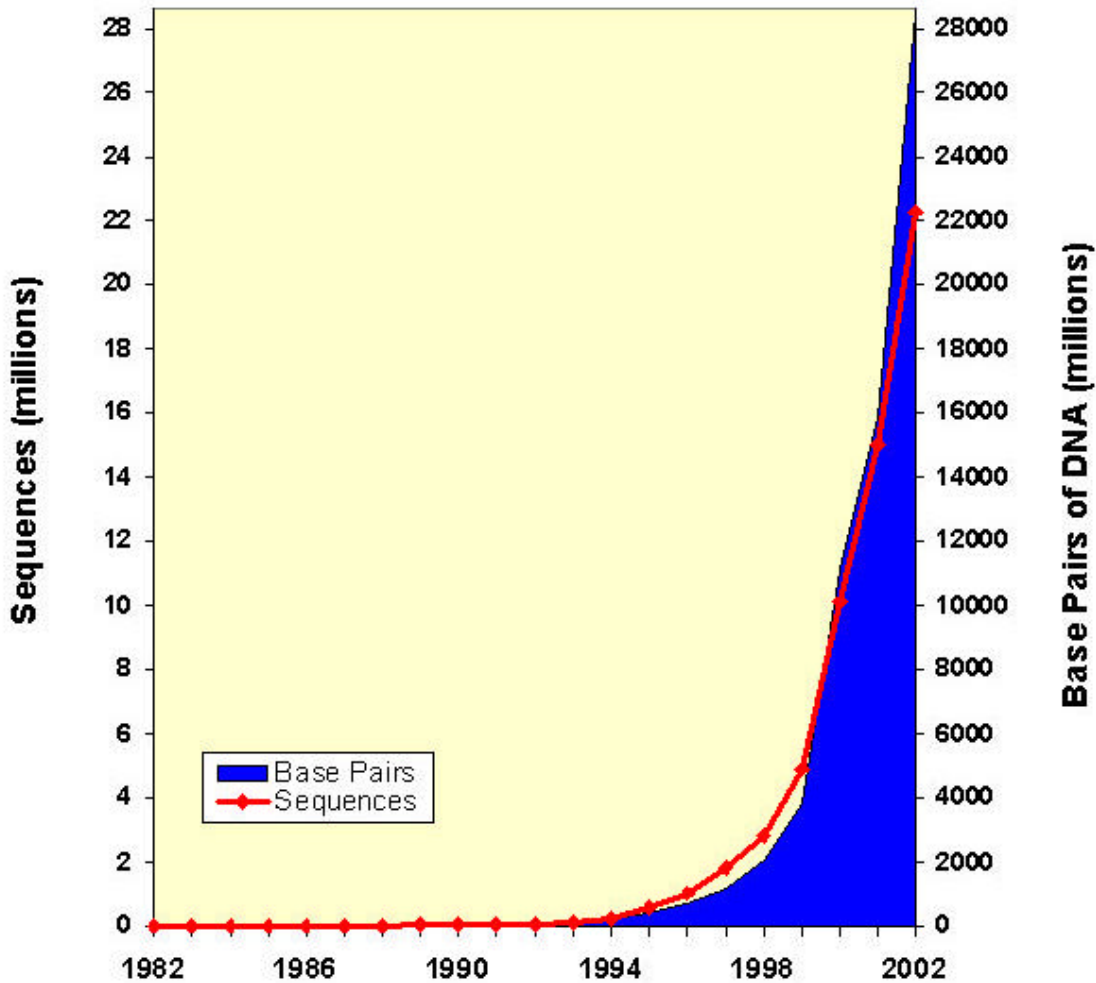


Figure 2-1 Growth of GenBank

Redundancies and database asynchrony are increasing because of the distributed and collaborative style of this research. As a result, database-to-database comparisons are required for analysis and validation, consuming ever more compute cycles and storage. The rate of acquisition of human and other genomic data over the next few years will be approximately 100 times higher than originally anticipated due to improved sequencing technology and methods. As complete genomes are sequenced, the length of DNA comparison strings will change from

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

single genes to entire genomes, with a concomitant expansion in the time to compute. To look at long-range patterns of expression synthetic regions on the order of 10's of megabases become reasonable lengths for consideration<sup>6</sup>. The challenges include: access, storage, and archiving.

## **2.2.Sequence and Alignment Software and Algorithms**

Biological R&D often requires the comparison of several sequences. Similarity comparisons evaluate the “closeness” of sequences to each other by computing a metric that includes a reward for allowed differences and penalty for disallowed differences. An objective function determines what rewards and penalties are important and how to combine these into the closeness metric. Corresponding to the modes of biology there are two types of sequence assessment: homology evaluation and contextual analysis. Both types of analyses and objective function are used to determine the best alignment of the sequences in question.

Homology evaluation looks for evidence that biological sequences are related by evolution. Orthologs are related molecules that have been changed due to speciation, while paralogs are replicated molecules in the same organism that have been altered through generations of independent mutation. Homology analyses depend on a proper analysis of related sequences because it is necessary to predetermine notions of which mutations are allowed and the rate they can be expected to occur.

Contextual analyses are used to join together many small sequences into fewer, longer sequences. They can also be used to find vector contamination in sequences, evaluate primer candidates, and do biochip design. Contextual analyses look for the common features among sequences without concern for whether the sequences have a common ancestor. The comparisons determine whether sequences overlap or are contained within another sequence<sup>7</sup>.

---

<sup>6</sup> <http://cbcg.llbl.gov/ssi-csb/Chapter2.html>

<sup>7</sup> <http://www.paracel.com/products/index.html>

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)

## **2.3. Sequence Alignment**

The objective of a sequence alignment algorithm is to position amino acid sequences so that the matched stretches of amino acids correspond to common structural or functional features. Gaps in the aligned sequences correspond to regions where polypeptide loops are deleted or inserted. Sequence alignment is a key component of many procedures for predicting the structure of a new protein whose sequence has just been determined. There are three general types of sequence-alignment method:

- Algorithms that attempt to match two sequences along their entire length
- Algorithms that search for local alignments involving sections (not necessarily continuous) from the sequences. The best known of these are Needle-Waterman and Smith and Waterman.
- Heuristic methods – BLAST and FASTA

A discussion of some of the most popular sequence-alignment applications and algorithms follows.

### **2.3.1. Basic Local Alignment Search Tool (BLAST) (many versions available)**

BLAST is a general purpose similarity search tool that may be used in contextual and homology analyses. It has good sensitivity and very good specificity and can report multiple local alignments between sequences. The BLAST algorithm is a heuristic search method. The programs use the statistical methods of Karlin and Altschul<sup>8</sup>. For a detailed description of the BLAST algorithm see [http://www.blc.arizona.edu/courses/bioinformatics/book\\_pages/blast.html](http://www.blc.arizona.edu/courses/bioinformatics/book_pages/blast.html). A public domain version of BLAST is available from the Blast server at <http://www.ncbi.nlm.nih.gov/BLAST/>. There are many variants of BLAST, including:

1. **BLASTN** - Compares a DNA query to a DNA database. Searches both strands automatically. It is optimized for speed, rather than sensitivity.

---

<sup>8</sup> S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403-410, 1990

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)

2. **BLASTP** - Compares a protein query to a protein database.
3. **BLASTX** - Compares a DNA query to a protein database, by translating the query sequence in the 6 possible frames, and comparing each against the database (3 reading frames from each strand of the DNA) searching.
4. **TBLASTN** - Compares a protein query to a DNA database, in the 6 possible frames of the database.
5. **TBLASTX** - Compares the protein encoded in a DNA query to the protein encoded in a DNA database, in the possible frames of both query and database sequences (Note that all the combinations of frames may have different scores).
6. **BLAST2** - Also called *advanced BLAST*. It can perform gapped alignments.
7. **PSI-BLAST** - (Position Specific Iterated) Performs iterative database using a traditional pairwise alignment algorithm<sup>9</sup> based on the Pearson and Lipman method.

### 2.3.1 FastA

FastA compares a DNA sequence to a DNA database or a protein sequence to a protein database.

Practically, FastA is a family of programs, which include: FastA, TFastA, Ssearch, etc.

<http://www2.ebi.ac.uk/fasta3/>. For a sketch of the algorithm see

<http://www.math.tau.ac.il/~rshamir/algmb/98/scribe/html/lec04/node14.html>.

### 2.3.2. Dynamic Programming and the Needleman-Wunsch Algorithm

The Needleman-Wunsch Algorithm is widely used for aligning pairs of sequences. The algorithm finds the optimal alignment based upon the scoring matrix used. [Needleman and Wunsch 1970]. The algorithm uses dynamic programming, which forms the basis for a number of widely used methods in bioinformatics. As mentioned in the introduction, sequence alignment is a ‘hard’ problem because there are an extremely large number of possible solutions, on the order of  $10^{30}$  for two sequences of length 100.

---

<sup>9</sup> R. W. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444-2448, 1988.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

### **2.3.3. Smith-Waterman (many variants available)**

The Smith-Waterman algorithm finds optimal, local alignment of nucleotide or peptide sequences and is typically used when low to moderate sequence identity is expected.

Alignments are optimal because the algorithm considers all possible ways that two sequences can be matched up and reports the one with the best score. The Smith-Waterman algorithm is a database search algorithm based on the Needleman and Wunsch algorithm. The Smith-Waterman algorithm uses dynamic programming to take alignments of any length, at any location, in any sequence, and determines whether an optimal alignment can be found. Based on these calculations, scores or weights are assigned to each character-to-character comparison: positive for exact matches/substitutions, negative for insertions/deletions. In weight matrices, scores are added together and the highest scoring alignment is output. Smith-Waterman is superior to the BLAST and FASTA algorithms because it searches a larger field of possibilities, making it more sensitive; however, individual pair-wise comparisons between letters slows the process down significantly.

Instead of looking at an entire sequence at once, the Smith-Waterman algorithm compares multi-length segments, looking for whichever segment maximizes the scoring measure. The algorithm itself is recursive.<sup>10</sup>

$$H_{i,j} = \max\{H_{i-1,j-1} + s(a_i, b_j); H_{i-k,j} - W_k; H_{i,j-1} - W_1; 0\}$$

Many groups use special hardware and software, such as Bioccelerator, to execute the algorithm.

See [http://dapsas1.weizmann.ac.il/bcd/bcd\\_parent/bcd\\_bioccel/bioccel.html](http://dapsas1.weizmann.ac.il/bcd/bcd_parent/bcd_bioccel/bioccel.html) .

---

<sup>10</sup> [http://www-cse.stanford.edu/classes/sophomore-college/projects-00/computers-and-the-hgp/smith\\_waterman.html](http://www-cse.stanford.edu/classes/sophomore-college/projects-00/computers-and-the-hgp/smith_waterman.html)

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)

### **2.3.4. Hidden Markov Models**

Hidden Markov Models (HMMs) are commonly used to specify protein profiles<sup>11</sup>. HMMs are built upon finite state machines with probabilities attached, i.e., stochastic regular grammars. HMMs have been generalized to recognize RNA secondary structure motifs using HMM algorithms with *stochastic context-free grammars* (SCFG) to capture conserved base-pairing. HMMs use position-specific scoring for the matching or substitution of a residue and for the opening or extension of a gap. HMMs are available from large, well-maintained libraries. HMMs have successfully been used in speech recognition as well as biology<sup>12</sup>. There are many variants available, see:

- HMMER is Sean Eddy's popular software for running HMMs. <http://hummer.wustl.edu>
- SAM (Sequence Alignment and Modeling Systems) is HMM software developed by Richard Hughey, Kevin Karplus and David Haussler at UC Santa Cruz<sup>13</sup>  
<http://cse.ucsc.edu/compbio/HMM-applicationsapps/HMM-.html>
- HMMpro is commercial HMM software developed by Piere Baldi and Yves Chavin at NetID Inc. <http://www.netid.com>.

### **2.3.5. CLUSTAL W**

CLUSTAL is one of the most popular packages for multiple sequence alignment. Multiple sequence alignment of nucleotide or protein sequences is an important tool in modern biology that helps reveal similarities or differences between various sequences. Its main features include carrying out multiple alignments of a large number of sequences with additional features for profile alignments (alignments of old alignments) and phylogenetic analysis. (Neighbor Joining

---

<sup>11</sup> R. Hughey and A. Krogh. Hidden Markov models for sequence for analysis:extension and analysis of the basic method. *Computer Applications in the Biosciences*, 12:95-107. 1996.

<sup>12</sup> Rabiner, L.R. "A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition", *Proceedings of the IEEE*, Vol. 77, pp. 257-286, February 1989.

<sup>13</sup> Kevin Karplus, Christian Barrett, Richard Hughey, "[Hidden Markov Models for Detecting Remote Protein Homologies](http://www.cse.ucsc.edu/research/compbio)", *Bioinformatics* 14(10):846-856, 1998. WWW server available from <http://www.cse.ucsc.edu/research/compbio>.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

trees can be calculated after multiple alignment with a bootstrapping option). The CLUSTAL alignment algorithm consists of 3 steps:

- calculation of pairwise sequence similarities in order to calculate a distance matrix giving a divergence of each pair of sequences
- construction of a guide tree (or a dendrogram) from the distance matrix
- multiple alignment of the sequences in a pairwise manner according to the branching order in the guide tree.

### **2.3.6. NONMEM**

NONMEM ("Nonlinear Mixed Effects Model") is a program for performing analyses of population pharmacokinetic/pharmacodynamic (PK/PD) data, written and distributed by the [NONMEM Project Group](#) at the [University of California at San Francisco](#). NONMEM uses general nonlinear regression techniques to fit models to data, in particular data collected from clinical drug studies.

### **2.3.7. PHYLIP**

PHYLIP (PHYLogeny Inference Package) is a package of programs for inferring phylogenies (evolutionary trees). It is maintained and developed by Dr. Joe Felsenstein ([University of Washington](#)). Methods that are available in the PHYLIP package include DNA and protein parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees<sup>14</sup>.

### **2.3.8. FastME**

FastME is a fast phylogeny reconstruction program based on the minimum evolution method. Among distance methods, FastME has shown better topological accuracy than Neighbor Joining, BIONJ, WEIGHBOR and FITCH. FastME first builds an initial tree, using either GME or BME algorithms, and then improves this tree by tree swapping, using either FASTNNI or BNNI

---

<sup>14</sup> Felsenstein, J. (1989). PHYLIP - Phylogeny inference package (version 3.2). *Cladistics*. 5: 164-6.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

algorithms. GME and FASTNNI optimize the ordinary least-squares (OLS) version of the minimum-evolution principle, while BME and BNNI optimize the balanced version<sup>15</sup>. A public version is available at: <http://www.ncbi.nlm.nih.gov/CBBresearch/Desper/FastME.html>

## **2.4. Bio Information Processing General Computing Overview**

Most genomics text searching algorithms are “embarrassingly parallel”; they can be deconstructed into a large number of independent searches with little message passing between jobs. When the independent jobs are completed, the final results are assembled. Computation is integer based. The computer systems used for the computation can be either SMPs, clusters of single CPU systems, or SMP clusters.

There is significant research into new kinds of statistical models. Hybrids of HMMs and neural nets, dynamic Bayesian nets, factorial HMMs, Boltzmann trees and hidden Markov random fields are among the areas being explored.<sup>16</sup> Stochastic grammars can be applied to biological sequences. SCFGs, in particular, and the corresponding learning algorithms have been used to derive statistical models of tRNA. However SCFGs have some limitations. First they are computationally intensive, so that in their present form they become somewhat impractical for long sequences, typically above N=200. Second not all RNA structures can be captured by an SCFG. The associated parse trees cannot capture tertiary interactions such as pseudoknots and non-pairwise interactions. Third they do not include a model for introns that present in some tRNA genes. Future requirements include<sup>17</sup>:

- Algorithmic and hardware speed improvements;

---

<sup>15</sup> Desper, R., Gascuel, O. (2002). Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum Evolution Principle, Proceedings of the 2<sup>nd</sup> Workshop on Algorithms in Bioinformatics (WABI), Roma, Lecture Notes in Computer Science.

A longer version (including algorithmic details and mathematical proofs) will be published in Journal of Computational Biology 19(4), 2002.

<sup>16</sup> R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids. Pp. 51-68. Cambridge University Press. 1998.

<sup>17</sup> Baldi, Pierre and Brunak, Soren. Bioinformatics The Machine Learning Approach. Pg. 297. MIT Press 2001.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

- Development of grammars, perhaps graph grammars, or other models, and the corresponding training algorithm to incorporate RNA tertiary structures, and possibly the tertiary structure of other molecules;
- Combination of SCFGs in modular ways, as for HMMs, to model more complex RNA sequences, including the corresponding introns;
- Modeling larger and more challenging RNA sequences, such as rRNA;
- Developing hybrid SCFG/NN architectures (or SG/NN), where NN is used to compute the parameters of a SCFG and/or to modulate or mix different SCFGs.

The following table provides current and expected capability requirements for major genomic codes:

Current and Expected Capability Requirements for Major Community Genomic Codes<sup>18</sup>

<b>Problem Class</b>	<b>Sustained Capability 1999</b>	<b>Sustained Capability 2002</b>
Sequence Assembly	$>10^{12}$ flops	$> 10^{14}$ flops
Binary Sequence Comparison	$>10^{12}$ flops	$> 10^{14}$ flops
Multiple Sequence Comparison	$>10^{12}$ flops	$> 10^{14}$ flops
Gene Modeling	$>10^{15}$ flops	$> 10^{17}$ flops
Phylogeny Trees	$>10^{11}$ flops	$> 10^{13}$ flops
Protein Family Classification	$>10^{10}$ flops	$> 10^{12}$ flops

<sup>18</sup> <http://cbcg.lbl.gov/ssi-csb/Meso.html>

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

### **3 Computational Biology**

Computational Biology includes molecular modeling, tissue engineering, organ modeling and systems biology. It is the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral and social systems.

#### **3.1 Molecular Modeling**

Molecular modeling includes biochemical analysis, protein binding/drug target evaluation, and dynamics of molecules. Molecular modeling techniques are widely used in the chemical, pharmaceutical and agrochemical industries. Molecular modeling techniques allow the simulation of systems of variable size, ranging from a few tens to millions of atoms. The parameters which rule the reliability of the simulation reside on the accuracy in the definition of the inter-atomic potential and on the dimensions of the investigated system, since a higher accuracy usually corresponds to increased computational requirement which in turn limits the dimension of the system under study.

A variety of modeling techniques have been developed over the years including:

- Quantum mechanical methods
- Density function theory
- Molecular Mechanics
- Energy minimization
- Molecular dynamics
- Monte Carlo methods
- Conformational analysis

##### **3.1.1 Quantum Mechanical Methods**

Quantum mechanical (QM) methods are used to determine energy interaction potentials. QM methods deal with the electrons in a system, so that a large number particles must be considered and the calculations are time-consuming. Quantum mechanics explicitly represents the electrons

**Federation of American Scientists**  
**Biomedical Computing Requirements for HPCS**  
**Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

in a calculation, and so it is possible to derive properties that depend upon the electronic distribution and to investigate chemical reaction in which bonds are broken and formed. There are two major categories of quantum mechanical molecular orbital calculations: ab initio and semi-empirical methods. The ab initio method uses the full Hartree-Fock/Roothaan-Hall equations, without ignoring or approximating any of the integrals or any of the terms in the Hamiltonian. Semi-empirical methods simplify the calculations, using parameters for some of the integrals and/or ignoring some of the terms in the Hamiltonian. Many different programs are available for performing ab initio calculations, the best known of these is the Gaussian series of programs.

An ab initio calculation [35, 36, 42, 48, 33] can be logically considered to involve two separate stages. First, the one- and two-integrals are calculated. This is computationally intensive. In the second stage, the wavefunction is determined using the variation theorem. In a traditional Self-consistent Field (SCF) calculation all of the integrals are first calculated and stored on disk, to be retrieved later during the SCF calculation as required. The number of integrals to be stored may run into millions and this leads to delays in accessing the data. In direct SCF calculation, the integrals are not stored on the disk but are kept in memory or recalculated when required<sup>19</sup>

Ab initio methods represent the higher level of description of the inter-atomic potential and allow, in principle, the exact solution of the Schrödinger equation without the introduction of any parameters. However, they usually offer a particularly unfavorable scaling with the dimensions of the system,  $O(M^8)$ , typical of many-body problems, which makes them applicable to systems composed by a limited number of atoms, usually of the order of 10–20. However, they can be extremely accurate, up to 0.5 kcal/mol, and are still successfully applied in the field of atmospheric chemistry and elementary chemical processes, in which high accuracies are needed. From a computational point of view, the most intensive tasks are represented by the analytic or numerical evaluation of 2-electron integrals, and integral transformations from an atomic orbital to a molecular orbital base. Conventional algorithms require the storage on disk

---

<sup>19</sup> Almlöf J, K Faegri and K Korsell 1982. Principles for a Direct SCF Approach to LCAO-MO *Ab initio Calculations*. Journal of Computational Chemistry 3: 386-399.

**Federation of American Scientists**  
**Biomedical Computing Requirements for HPCS**  
**Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

of an enormous amount of data, the semi-transformed integrals, of the order of tens of Gbytes, as well as large memory storage, bandwidth and latency. This class of applications can be defined as memory-bound and is indeed bound to the efficiency of the memory access on a single processor; moreover, the extremely involved data connectivity, makes these algorithms difficult to implement on parallel machines requiring more and more efficient single CPUs.

### **3.1.2 Density Functional Theory**

Density Functional Theory (DFT) attempts to calculate the total electronic energy and the overall electronic density distribution. DFT methods usually offer an  $O(M^3)$  scaling with the dimensions of the system, typical of direct diagonalization techniques. More favourable scaling, of the order of  $M^2 \log M$ , can be achieved by using a plane wave (PW) expansion as a basis set and pseudo-potentials (PP) for the description of core electrons<sup>20</sup>. The typical accuracy of these methods, of the order of 3-7 kcal/mol, makes them suitable to the study of chemical interesting problems, and DFT methods have been successfully applied to the investigation of chemical reactivity and complex material in systems composed by up to a few hundreds of atoms. Moreover, recent developments [24], include coupling the evaluation of a DFT potential to a classical molecular dynamics (MD) scheme, introducing time as a further degree of freedom to explore. The basic algorithmic features of DFT-based MD methods reside in the use of efficient fast Fourier transform (FFT) techniques to compute the different contributions to the total energy (kinetic energy, Coulomb, XC, PP) and its derivatives, the latter task being particularly computationally intensive. The large number of PWs, typically of the order of  $10^5$ - $10^6$ , necessarily translates in large memory requirements; moreover, the parallel implementation of this class of algorithms requires an extremely efficient communication network, due to the particular implementation of the parallel FFT which requires global data exchange. A system composed by 350 atoms can require up to 24 Gbyte of memory<sup>21</sup>. This class of methods will therefore benefit from both increased computer power, communication and memory bandwidth, even if it will be probably limited to run on proprietary hardware, due to the reduced performances of COTS

---

<sup>20</sup> <http://www.fyslab.hut.fi/epm/abinitio/pwpp/overview/>

<sup>21</sup> <http://www.epcc.ed.ac.uk/enacts/hpcroadmap.pdf>

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

communication devices. Moreover, an adequate development of scientific libraries (FFT and linear algebra) is needed to retain high performance for this class of algorithms.

### **3.1.3 Molecular Mechanics**

Molecular mechanics, also known as force field methods, are used to perform calculations on systems containing significant numbers of atoms. Force field methods ignore the electron motions (the focus of quantum mechanics) and calculate the energy of a system as a function of the nuclear positions only. In some cases force fields can provide answers that are as accurate as even the highest-level quantum mechanical fraction of the computer time. Molecular mechanics is based on a simple model of the interactions within a system with contributions from processes such as the stretching of bonds, the opening and closing of angles and the rotations about single bonds.

The interaction potential is usually expressed as the sum of a priori parameterised van der Waals and Coulombic contributions, the latter showing a quadratic scaling,  $O(M^2)$ , with the dimensions of the system (generally indicated with  $M$ ), due to the double sum on the atomic effective charges. The maximum accuracy of this class of methods is typically of the order of 20 kcal/mol, usually too limited for the description of phenomena of chemical interest. However, FF-based methods, implemented according to the fast-multipole (FM) expansion of the Coulomb potential, have been applied to the approximate description of systems containing up to a few million of atoms and have found a wide success in the investigation of biological systems, surface science and material science (e.g. protein science, material fractures, liquid crystals). FF-based algorithms usually offer excellent scaling performances on massively parallel architectures. FM expansion of the Coulomb potential can be implemented in such a way as to reach a linear scaling with the dimensions of the system [22].

A study by the European Network for Advanced Computing Technology for Science (HPC Technology Roadmap) projected that in within the next years, DFT methods will probably allow

**Federation of American Scientists**  
**Biomedical Computing Requirements for HPCS**  
**Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

the accurate computations of electronic, structural and dynamical reactive properties of systems containing 10000 atoms. Based on this, they predict that DFT methods will substitute FF parameterizations in chemiometric applications, in which a large number of medium-size calculations is needed. They predict that this will have a direct impact in pharmacology: “the design of a new drug usually requires a pre-selection operated by computer simulations and data analysis; the advantage of a much higher accuracy in the description of the investigated molecular systems and properties directly translates into a high selectivity of the target system with a significant reduction of the number of laboratory tests, up to a factor of 10. DFT methods will also allow the accurate simulation of small protein systems, or of realistic portions of them, with particular impact on the comprehension of the action mechanism of metalloenzymes, where a reduced model usually neglects the fundamental underlying interactions. To understand the importance of such a field, it is sufficient to mention that both respiration and photosynthesis involve metallo-organic active centres constituted by several thousand atoms; comprehension of the action mechanism of such systems will allow to device efficient synthetic bio-mimetic analogues of the natural systems, with a high impact in the field of energy storage and molecular sensors. Moreover, we can predict that DFT-based methods will allow the accurate simulation of nano-scale systems with a high impact in the design of molecular engines, quantum computation devices and chemical storage of data<sup>22</sup>.”

### **3.1.4 Energy Minimization**

Energy minimization is widely used in molecular modeling and is an integral part of techniques such as conformational search procedures. Minimum energy arrangements of the atoms correspond to stable states of a system. Energy minimization is also used to prepare a system for other types of calculations, for example it may be used prior to a MD or Monte Carlo simulation. Molecular mechanics minimizations are nearly always performed in Cartesian coordinates where the energy is a function of  $3N$  variables. Minima are located using numerical methods that gradually change the coordinates to produce configurations with lower and lower energies until the minimum is obtained. Algorithms used include the simplex or steepest descents methods and

---

<sup>22</sup> C. W. Bauschlicher, A. Ricca, and R. Merkle. Chemical storage of data. *Nanotechnology*, 8:1, 1997.

**Federation of American Scientists**  
**Biomedical Computing Requirements for HPCS**  
**Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

the Newton-Raphson algorithm. Systems containing thousands of atoms can require significant memory, and are usually solved using molecular mechanics methods.

### **3.1.5 Molecular Dynamics**

Molecular dynamics (MD) calculates the real dynamics of the system from which time averages or properties can be calculated. Sets of atomic positions are determined in sequence by applying Newton's equations of motion. A MD simulation generates a trajectory that describes how the dynamic variables change with time. MD simulations typically run for tens of hundreds of picoseconds. Thermodynamic averages are obtained from MD as time averages using numerical integration of the following equation:

M is the number of time steps. MD is also used extensively to investigate the conformational properties of flexible molecules.

### **3.1.6 Monte Carlo Methods**

In a Monte Carlo simulation the outcome of each trial move depends only upon its immediate predecessor, whereas in molecular dynamics it is possible to predict the configuration of the system at any time. In a Monte Carlo simulation the total energy is determined directly from the potential energy function. Monte Carlo simulation samples from the canonical ensemble (constant N, V and temperature, T). Monte Carlo methods use a technique called importance sampling, which are able to generate states of low energy. The Metropolis method is used to generate configurations that make a large contribution to the integral. The Metropolis method is derived by imposing the condition of microscopic reversibility: at equilibrium the transition between two states occurs at the same rate. The rate of transition from state  $m$  to a state  $n$  equals the product of the population and the appropriate element of the transition matrix.

MC and MD complement each other in their ability to explore phase space. MC often give more rapid convergence of the calculated thermodynamic properties of a simple molecular liquid, but it may explore the phase space of large molecules very slowly due to the need for small steps

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

unless special techniques such as the configurational bias MC methods are employed. However, the ability of the MC methods to make non-physical moves can significantly enhance its capacity to explore phase space in where there are a number of minimum energy states separated by high barriers. MD may not be able to cross the barriers between the conformations sufficiently often to ensure that each conformation is sampled according to the correct statistical weight. MD advanced the positions and velocities of all the particles simultaneously and it can be very useful for exploration of the local phase space whereas the MC methods may be more effective for conformational changes, which jump to completely different area of phase space. A variety of hybrid MD/MC methods have been used, in which the simulation algorithm alternates between MD and MC. The aim of such methods is to achieve better sampling, and thereby more rapid convergence of thermodynamics properties. Such a hybrid systems has been used to perform long simulations of DNA molecules.<sup>23</sup>

### **3.1.7 Conformational Analysis**

The physical, chemical and biological properties of a molecule depend upon the 3D structures or conformations that it can adopt. Conformational analysis is the study of the conformations of a molecule and their influence on its properties.<sup>24</sup> A key component of a conformational analysis is the conformational search, which seeks to identify the preferred conformations of a molecule; this requires locating conformations that are a minimum points on the energy surface. Thus, energy minimization methods are crucial in conformational analysis. It is desirable to identify all minimum energy conformations on the energy surface, however the number of minima may be so large that it is impractical to contemplate finding them all. Under these circumstances it is often assumed that the naturally occurring conformation is the one with the lowest minimum energy function. The conformation is usually referred to as the global minimum energy conformation. Conformational search methods can be divided into the following categories:

- Search algorithms
- Model-building methods

---

<sup>23</sup> Seaminathan S., G. Ravishanker and D. L. Beveridge 1991. Molecular Dynamics of B-DNA Including Water and Counterions – A 140-ps Trajectory for d(CGCGAATTCGCG) Based on the Gromos Force Field. *Journal of the American Chemical Society*. 113:5029-5040.

<sup>24</sup> Leach, Andrew...

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

- Random approaches
- Distance geometry

### **3.1.7.1 Search Algorithms**

A systematic search explores the conformational space by making regular and predictable changes to conformation. The simplest type of systematic search, often called a grid search. All rotatable bonds in the molecule are identified. The lengths and angles remain fixed throughout the calculation. Each of these bonds is systematically rotated through 260 degree using a fixed increment. Every conformation so generated is subjected to energy minimization to derive the associated minimum energy conformation. The search stops when all possible combinations of torsion angles have been generated and minimized.

To partially alleviate the combinatorial explosion that accompanies a systematic search is to use larger building blocks, or molecular fragments to construct the conformations. Model-building approaches construct conformations of a molecule by joining together 3-D structures of molecular fragments. A conformation of the molecule is constructed by assigning a template to each fragment and then attempting to join the templates together. The search problem can be represented as a tree, as for a systematic search, and so all of the usual tree-searching algorithms are applicable.

### **3.1.7.2 Random Search Methods**

Random search methods explore conformational space by changing either the atomic Cartesian coordinates or the torsion angles of rotatable bonds. Both algorithms use a similar approach. At each iteration, a random change is made to the current conformation. The new structure is then refined using energy minimization. If the minimized conformation has not been found previously, it is stored. The conformation to be uses as the starting point for the next iteration is then chosen and the cycle starts again. The procedure continues until a given number of iterations have been performed or until it is decided that no new conformations can be found.

### **3.1.7.3 Distance Geometry**

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

The conformation of a molecule can be described in terms of the distances between all pairs of atoms. There are  $N(N-1)/2$  interatomic distances in a molecule, which are most conveniently represented using an  $N \times N$  symmetric matrix. In the matrix, the elements  $(i,j)$  and  $(j,i)$  contain the distance between atoms  $i$  and  $j$  and the diagonal elements are all zero. Distance geometry explores conformational space by randomly generating many distance matrices, which are then converted into conformations in Cartesian space. The crucial feature is that it is not possible to arbitrarily assign values to the interatomic distance in a molecule and always obtain a low-energy conformation. The interatomic distances are closely interrelated and many combinations of distances are geometrically impossible. Distance geometry uses a four-stage process. First, a matrix of upper and lower interatomic distance bounds is calculated. This matrix contains the maximum and minimum values permitted to each interatomic distance in the molecule. Values are then randomly assigned to each interatomic distance between its upper and lower bounds. In the third step, the distance matrix is converted into trial set of Cartesian coordinates. In the fourth step, the Cartesian coordinates are refined. See <http://www.scripps.edu/case/nab5/NAB-sh-5.1.html> for a description of the algorithm.

### **3.1.8 Evolutionary Algorithms and Simulated Annealing**

Evolutionary algorithms and simulated annealing have found widespread use in molecular modeling, include use in finding the global minimum energy conformation of a molecule, protein-ligand docking, molecular design, QSAR and pharmacophore mapping<sup>25</sup>. There are three basic classes of evolutionary algorithm:

- Genetic algorithms (GA)
- Evolutionary programming (EP), and
- Evolutional strategies (ES).

All three are based on the concept of creating a ‘population’ of possible solutions to the problem. The members of the population are scored using a ‘fitness function’ that measures how ‘good’ they are. The population changes over time and evolves towards better solutions. For a

---

<sup>25</sup> Clark D. E and D. R. Westhead 1996. Evolutionary Algorithms in Computer-Aided Molecular Design. *Journal of Computer-Aided Molecular Design*. 10:337-358.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

description of genetic algorithms,

<http://www.techfak.unibielefeld.de/bcd/Curric/ProtEn/contents.html>

### **3.1.8.1 Canonical Genetic Algorithm**

In the genetic algorithm<sup>26</sup>, a population of  $\mathcal{N}$  possible solutions is created. The population corresponds to a set of randomly generated conformations of the molecule, with each member of the population coded by a chromosome. This is typically stored as a linear string of bits. The chromosome codes for the values of the torsion angles of the rotatable bonds in the molecule. After decoding each chromosome and assigning the torsion angles to the appropriate values in the molecule, the fitness of each member of the population is calculated. An appropriate fitness function would be the internal energy, as might be calculated using molecular mechanics. A new population is then generated.  $\mathcal{N}/2$  pairs of parents are selected from the current population. A technique called roulette wheel selection is often used to achieve bias toward the most fit individuals but using slot sizes in the roulette wheel that are proportional to the values of the fitness function. The use of the roulette wheel selection means that particularly fit members of the population may be able to produce many offspring. The new population is then subjected to genetic operators, the two most commonly used of which are crossover, or recombination, and mutation. In crossover, a cross position  $i$  is randomly selected ( $1 \leq i \leq l$ , where  $l$  is the length of the chromosome). Two new strings are then created by swapping the bits between positions  $i + 1$  and  $l$ . The crossover operator is applied to the selected pairs of parents with a probability  $P_c$ , a typical value being 0.8. Following the crossover phase mutation is applied to all the individuals in the population. Each bit may be inverted (0 to 1 and vice versa) with a probability  $P_m$ . The new population then becomes the current population ready for a new cycle. The algorithm repeatedly applies this sequence for a predetermined number of iterations and or until it converges.

### **3.1.8.2 Evolutionary Programming**

---

<sup>26</sup> Goldberg, D.E 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, MA., Addison-Wesley.

**Federation of American Scientists**  
**Biomedical Computing Requirements for HPCS**  
**Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

The main difference between the genetic algorithm and evolutionary programming is that the latter does not use a crossover operator. Evolutionary strategies are very similar to evolutionary programming but differ in two key respects: crossover operators are permitted and the probabilistic tournament is replaced with a straightforward ranking. Genetic and evolutionary algorithms involve a significant random element and so they are not guaranteed to produce the same global minimum energy conformation from each run. They are useful for producing solutions very close to the global optimum in a reasonable amount of time. It is common practice to perform several runs in order to obtain a variety of different solutions and to investigate the nature of the energy surface.

### **3.1.8.3 Simulated Annealing**

Simulated annealing is a computational method that mimics annealing, the process in which the temperature of a molten substance is slowly reduced until the material crystallises to give a large single crystal. The perfect crystal that is eventually obtained corresponds to the global minimum of the free energy. Simulated annealing is used to find the optimal or best solutions to problems which have a large number of possible solutions. Simulated annealing is a general purpose optimization algorithm. It combines Markov-Chain Monte-Carlo methods (MCMC) ideas such as the Metropolis algorithm with a schedule for lowering temperature<sup>27</sup>.

In simulated annealing a cost function takes the role of the free energy in physical annealing and a control parameter corresponds to the temperature. To use simulated annealing in conformational analysis the cost function would be the internal energy. At a given temperature the system is allowed to reach ‘thermal equilibrium’ using a molecular dynamics or Monte Carlo simulation. At high temperatures the system is able to occupy high-energy regions of the conformational space and to pass over high energy barriers. As the temperature falls, the lower

---

<sup>27</sup> Baldi, Pierre and Soren Brunak. 2001. Bioinformatics: the machine learning approach. pg. 91. 2<sup>nd</sup> ed. The MIT Press.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

energy states become more probable in accordance with the Boltzmann distribution. At absolute zero, the system should occupy the lowest-energy state – the global minimum energy conformation. To guarantee that the globally optimal solution is reached would require an infinite number of temperature steps, at each of which the system would have to come to thermal equilibrium. Careful temperature control is required when the energy of the system is comparable with the height of the barriers that separate one region of conformational space from another. This is often difficult to achieve in practice and simulated annealing cannot guarantee to find the optimal solution. However, if the same answer is obtained from several different runs then there is a high probability that it corresponds to the true global minimum. Several simulated annealing runs may enable a series of low-energy conformations of a molecule to be obtained.

#### **3.1.8.4 Clustering Algorithms and Pattern Recognition Techniques**

Molecular modeling programs generate large quantities of data that must be processed and analyzed. Many conformations search algorithms can generate conformations that are very similar, if not identical. Cluster analysis is used to select from the data a smaller, representative set of conformations for subsequent analysis. A common use of cluster analysis is in selecting a set of representative molecules from a large chemical database.

A cluster analysis requires a measure of the similarity between pairs of objects. A large number of cluster algorithms are available. Hierarchical clustering involves a series of iterations at each which the two closest clusters are identified and combined into a larger cluster. These methods produce a clustering that is independent of the order in which the objects are stored. Simple implementations require an  $M \times M$  similarity matrix to be calculated, limiting their applicability when clustering large data sets. The Jarvis-Patrick method is a non-hierarchical clustering method that uses a nearest neighbors' approach. The algorithm can be used to cluster very large data sets. The K-means method is another non-hierarchical clustering method.

#### **3.1.9 Applications**

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)*

A number of popular software packages for molecular modeling are available and used widely, including:

- GAMESS programs for *Ab initio* Quantum Chemistry
- GAUSSIAN programs for *Ab initio* Quantum Chemistry [www.gaussian.com](http://www.gaussian.com)
- NWCHEM programs for Quantum Mechanics
- CHARM programs for molecular mechanics <http://yuri.harvard.edu>
- AMBER programs for molecular mechanical force field <http://amber.ucsf.edu>
- MOPAC/AMPAC programs for semi-empirical quantum mechanics
- MM2 program for molecular mechanics

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

## **4. Protein Biochemistry**

Protein Biochemistry includes protein structure and proteomics. It includes the identification, characterization and quantification of all proteins involved in a particular pathway, organelle, cell, tissue, organ or organism that can be studied in concert to provide accurate and comprehensive data about that system.

Determining the shape of proteins from their sequences is one of today's great computational challenges. The properties of any protein are largely determined by its structure. Proteins usually adopt a single structure, corresponding to the global minimum free energy under physiological conditions. Protein sequences can generally fold into a unique state in just a few seconds (or less) from any starting conformation. Protein structures can be experimentally determined by crystallizing the protein and then using x-ray crystallography or NMR to find the position of the atoms, but this is a difficult procedure. The experimental process of deciphering the atomic structures of the majority of cellular proteins is expected to take a century at the present rate of work.<sup>28</sup> Thus, the interest in using computational methods to predict protein structure. The folded structure of a sequence is determined by the sequence of successive solid bend angles, where each angle can be represented by two planar angles. It is possible to make such a problem discrete by limiting the ways to bend each angle, but doing so decreases the accuracy of the solution. Even with such techniques, a 100-residue protein would have a search space of  $7^{100}$  ( $\sim 10^{84}$  configurations).

---

<sup>28</sup> T. Head-Gordon and J.C. Wooley, IBM Systems Journal, Vo. 40, No 2, 2001, p. 265-293

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)

**The Levinthal Paradox** describes the discrepancy between the time for an exhaustive search of all possible conformations and the observed timescale of protein folding. If it is assumed that there are three conformations for each amino acid then a polypeptide chain with say 1— amino acids would have about  $10^{48}$  conformations. If the interconversion between conformations required just  $10^{-11}$  seconds then it would take about  $10^{29}$  years to explore them all. Of course, this is for the most basic of grid search algorithms, but even the most advanced systematic conformational search would still require an inordinate amount of time to identify the global minimum energy conformation.

Example: [Duan and Kollman 1998]. 1  $\mu$ s simulation of a 36-residue peptide starting from a fully extended state. This peptide is one of the smallest proteins that can fold autonomously, with folding estimated to take between 10  $\mu$ s and 100  $\mu$ s. It contains three short alpha-helices. The simulation involved in addition to the protein about 3000 water molecules and was performed in a truncated octahedron simulation box with a time step of 2 fs. About 4 months of computing time on a 256-processor parallel computer was required for the 1  $\mu$ s simulation. While the protein did not actually fold into the known experimental structure, a marginally stable state which showed significant resemblance to the native conformation was observed. This state had a lifetime of about 150ns.<sup>29</sup>

## 4.1 Protein Folding

A variety of approaches have been used for protein folding. The most ambitious approaches attempt to solve it *ab initio*. The conformational space of the molecule is explored to identify the appropriate structure. The total number of conformations is very large, and so it is usual to try to find only the very lowest energy structures. Some form of empirical force field is generally used, often augmented with a solvation term. The global minimum in the energy function is assumed to correspond to the naturally occurring structure of the molecule. Rule-based methods, often called threading, have also been used for protein folding. This approach first determines which stretches of amino acids should adopt each type of secondary structure and then packs these secondary structural elements together to achieve a low-energy structure. The threading approach relies on the quality of the initial secondary structure prediction. It works best if the structural class to which the protein belongs is known. A third approach, comparative modeling, exploits the structural similarities between proteins by constructing a 3-D structure based on the

---

<sup>29</sup> <http://www.psc.edu/science/Kollman98/kollman98.html>

**Federation of American Scientists**  
**Biomedical Computing Requirements for HPCS**  
**Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

known structure(s) of one or more related proteins. When using comparative modeling, one must initially determine which protein structure(s) to use as the 3D templates, and then decide how to match the amino acids in the unknown structure with the amino acids in the known structure(s). Each of these methods is described in the following sections.

#### **4.1.1 *Ab Initio* Prediction**

*Ab initio* prediction programs work by defining a global energy function and performing a search of possible bond-angle configurations to find one which minimizes total energy. *Ab initio* approaches explore the conformational space of the molecule to identify the appropriate structure. Since the total number of possible conformations is very large, it is usual to try to find only the lowest energy structures. Some form of empirical force field is usually used, often augmented with a solvation term.

Many methods are used for exploring the conformational space, many of which are analogous to the models used to perform Monte Carlo simulation of polymers, such as the lattice and ‘bead’ models<sup>30</sup>. An optimization procedure based on simulated annealing or a genetic algorithm is often used with simplified molecular dynamics models to first identify families of low-energy structures, which may then be converted into a more detailed representation for subsequent refinement<sup>31</sup>. The most important issues are:

- 1) the energy function selected – energy minimization functions include hydrophobic/hydrophilic interactions; size and flexibility properties of different amino acids; and electrostatic/Van der Waals interactions of nearby atoms;
- 2) the optimization procedure employed to search the space – methods include gradient descent, simulated annealing, and genetic algorithms, possibly using parallel computation.

---

<sup>30</sup> Sklonick J., A Kolinsik and A. R. Ortiz 1997. MONSSTER: A Method for Folding Globular Proteins with a Small Number of Distance Restraints. *Journal of Molecular Biology* 265:217-214.

<sup>31</sup> Leach, Andrew R. *Molecular Modeling Principals and Applications*. Prentice Hall 2001. p. 518.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)

One of the simplest and most popular biophysical models of protein folding is the hydrophobic-hydrophilic (HP) model<sup>32</sup>. The HP model abstracts the hydrophobic interaction in protein folding by labeling the amino acids as hydrophobic (H for nonpolar) or hydrophilic (P for polar). Chains of amino acids are configured as self-avoiding walks on the 3D cubic lattice, where an optimal conformation maximizes the number of adjacencies between H's. The protein folding problem under the HP model on the cubic lattice is shown to be NP-complete. For a discussion see <http://www.stanford.edu/class/cs374/Papers/p05-3.pdf>.

#### **4.1.2 Threading or Fold Assignment Approaches**

Many programs use known 3D structures to help determine a protein's 3D structure. Two amino acid sequences with 20% - 30% identical residues likely have similar 3D structures. Threading, or inverse folding, programs are commonly used<sup>33</sup>. The basic concept is to choose from among a number of 3D protein structures, typically chosen to represent a common structural class, choose the structure most compatible with the sequence of the unknown protein. This is accomplished by "threading" the sequence through each protein structure in turn. Threading methods are closely related to ad initio approaches to protein structure prediction, but threading methods inherently limit the search space to the conformations of known structures.

Threading programs use special searching methods such as double dynamic programming to efficiently find the best ways to match the sequence to the structure. Approximations are used to make the problem more manageable. Many of the scoring functions used in threading algorithms are potentials of mean force that provide an estimate of the free energy of interaction between two residues as a function of their separation. These potentials of mean force are calculated from statistical analyses of known protein structures. For threading algorithms one is particularly interested in the interactions between amino acids that are close in 3D space but far apart in the sequence, and the potentials used in such calculations are derived appropriately. In

---

<sup>32</sup> K- Dill, S. Bromberg, K. Yue, K. Fiebig, D. Yee, P. Thomas, and H. Ghan. Principles of protein folding. A perspective from simple exact models. *Protein Science*, 4561-602, 1995.

<sup>33</sup> Jones D.T. and J. Thornton 1993. Protein fold Recognition. *Journal of Computer-Aided Molecular Design* 7:439-456.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

addition, the pairwise knowledge-based term, a solvation contribution, is often added. Although knowledge-based potentials are most popular, it is also possible to use other types of potential function.

No one single theoretical or experimental technique can predict protein function from sequence, rather it is the application of an appropriate combination of methods that is required<sup>34</sup>. There are two main approaches:

- 1) developing potentials for fold assignment<sup>35,36</sup>
- 2) HMMs that are descended from alignment methods<sup>37,38</sup>

The input is 1) a protein structure, 2) a core model describing the position of the core residues and allowable lengths of loops and 3) a scoring function to evaluate the given threading.

Without modeling pairwise interactions this is a simple dynamic programming problem. It has been estimated that the success rate of fold assignment algorithms will increase to roughly 50% once the library of protein folds grows.<sup>39</sup> For the remaining genome sequences to be assigned to folds, it will be necessary to move to multi-positional compatibility functions. Incorporating pairwise interactions will require tabulating the possible substructures for every base assignment, not just the best matching prefix structures.

HMMs are used for fold identification by performing a standard sequence-based homology search using the probe sequence to generate homologous sequences. These sequences can be used to construct an HMM based on the probe, and then sequences from a library of folds can be matched against the HMM. HMMs can also be used to construct separate HMMs for each member of a library of folds and then score the probe sequence against each model. Construction

---

<sup>34</sup> Leach, Andrew R. *Molecular Modeling Principles and Applications*. Prentice Hall 2001. p. 549.

<sup>35</sup> R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge (1998).

<sup>36</sup> S.R. Eddy, "Profile Hidden Markov Models," *Bioinformatics* 14, 755-763 (1998).

<sup>37</sup> D. Sankoff and J.B. Kruskal, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison-Wesley Publishing Co., Reading, MA (1983).

<sup>38</sup> A. Sali and T.L. Blundell, "Definition of General Topological Equivalence in Protein Structures: A Procedure Involving Comparison of Properties and Relationships Through Simulated Annealing and Dynamic Programming," *Journal of Molecular Biology* 212, 403-428 (1990).

<sup>39</sup> T. Head-Gordon and J.C. Wooley, *IBM Systems Journal*, Vo. 40, No 2, 2001, p. 265-293

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

of HMMs is typically an iterative process involving successive periods of model building, searching with the given model, and model refinement. Alignment to an HMM can be performed in an efficient recursive manner, similar to dynamic programming.

### **4.1.3 Comparative Modeling**

Comparative modeling exploits the structural similarities between proteins by constructing a 3D structure based upon the known structures of one or more related proteins. To do this, it is necessary to decide which protein structures to use as 3D templates, and then to decide how to match the amino acids in the unknown structure with the amino acids in the known structures.

Comparative modeling methods consist of the following sequential steps:

- 1) identify the proteins with known 3D structures that are related to the target sequence
- 2) align these with the target sequence and pick those known structures that will be used as templates
- 3) build the model for the target sequence given its alignment with the template structures
- 4) evaluate the model against selected criteria
- 5) if necessary, repeat the alignment and model building until a satisfactory evaluation is reached.

In a typical comparative modeling exercise one would use a heuristic algorithm to determine possible sequences of interest, then the Smith-Waterman method to identify the appropriate subsequences, and finally the Needleman-Wunsch algorithm to derive the alignment to use in the actual construction of the model.

There are three different classes of method for constructing the 3D model. Generally, each of these three methods is used, with construction proceeding as in the following three-stage process.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

- Piece together rigid bodies taken from the template protein(s). This step constructs the model from a few core regions, loops and side chains obtained from dissected related structures<sup>40</sup>.
- Assemble the target protein by joining together small segments or by reconstructing a set of coordinates. Segment matching relies on approximate positions of conserved atoms from the templates to calculate the coordinates of other atoms; this is achieved by the use of a database of short segments or protein structure, energy or geometry rules, or some combination of these criteria. Fragment assembly without using an underlying framework. The fragments are taken from proteins of known structure which show local sequence similarity to the unknown target. The initial structures resulting from this “splicing” process are then subjected to simulated annealing using a scoring function that has sequence-dependent terms and sequence independent terms. The most promising of several runs are selected<sup>41</sup>.
- Generate a series of spatial restraints from the templates, which are used in conjunction with an optimization procedure to derive a structure of the target. Satisfaction of spatial constraints uses either distance geometry or optimization techniques to satisfy spatial restraints obtained from the alignment of the target sequence with homologous templates of known structure. The optimization uses a combination of conjugate gradients, with molecular dynamics and simulated annealing.

---

<sup>40</sup> Srinivasan N., K. Gurpasad and T. L. Blundell 1996. Comparative Modeling Proteins. In Sternberg M E (Editor) *Protein Structure Prediction – A Practical Approach*. Oxford, IRL Press, pp. 111-140.

<sup>41</sup> Simons K. T., R. C Kooperberg, E Huang and D Baker 1999b. Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences Using Simulated Annealing and Bayesian Scoring Functions. *Journal of Molecular Biology* 268:209-225.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)

## 5.0 Drug Discovery

Discovering and developing any new medicine is a long and expensive process. A new compound must not only produce the desired result with minimal side-effects but must also be demonstrably better than existing therapies. Typically, the two key steps in drug discovery programs are the identification of “hit” molecules and lead series, or “leads”. A *hit* is a molecule that has some reproducible activity in a biological assay. A *lead series* comprises a set of related molecules that usually share some common structural feature and which show some variation in the activity as the structure is modified. This provides confidence that further synthetic modification to the lead series has a good chance of resulting in a drug candidate with the desired potency and selectivity, lack of toxicity and appropriate characteristics to enable it to reach its target in vivo. Such a drug candidate will then enter the early stages of development, where further large-scale investigations are undertaken.

Although high-throughput screening makes it possible in principle to test every available compound against every biological assay, there are number of practical reasons why this is not feasible:

- The large number of samples now available in many companies means that the overall expense can be significant
- Some assays cannot be converted to a high-throughput format and so have to be conducted using more traditional technique
- A significant proportion of the available samples might not be considered appropriate structures.

As a result, it is often necessary to identify subsets of compounds. Computational techniques play a significant role in determining which such subsets can be constructed, with various techniques being available depending upon the type of molecule to be screened, the information available to assist with the selection and the properties to be taken into account.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

A wide variety of methods are used either individually or in combination to select compounds. 2-D methods use only information about the chemical structure of the molecule. 3-D methods use information about the molecules confirmation and properties dependent upon the confirmation. Some methods take into account information about the target protein or about other molecules that are known to be active at the target, whereas other methods are designed to produce diverse collections of compounds for more general screening.

Having tested a number of compounds, a model is usually constructed that relates the observed activity to the molecular structure. The model can then be used in the next iteration of the process. Many different kinds of models are used. A popular approach is to use statistical techniques to derive the model.

### **5.1 Substructure Searching**

Substructure searching is the most basic approach to identifying compounds of interest. Many organizations maintain databases of chemical compounds; some are non-proprietary and some are proprietary. A database may consist of large numbers of compounds, several hundred thousand is common. The American Chemical Society database contains more than 18 million compounds.<sup>42</sup> Most systems represent molecules as molecular graphs. A graph contains nodes, which are connected by edges. A subgraph is a subset of the nodes and edges of a graph. A key requirement for any chemical database system is that it can determine whether or not a new molecule is already present in the systems. A substructure search retrieves all the molecules from the database that contain the substructure. Substructure searching is known as subgraph isomerism – determining whether one graph is entirely contained within another. Even with the most efficient algorithms this is a relatively time-consuming process and so chemical database systems commonly use some form of screening method to rapidly eliminate molecules that cannot match the query. Such screens are frequently implemented using binary representations and so operate rapidly, especially if held in memory.

---

<sup>42</sup> Leach pg. 642.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

### **5.1.1 Binary Screening**

Two types of binary screening are used. In a *structural key*, each position in the bitstring corresponds to a particular substructure. If that substructure is present in the molecule, then the relevant bit in the molecule's key is set to 1. A predefined fragment dictionary is used to specify the substructures. As each molecule is added to the database a substructure search is performed for each fragment and the relevant bit assigned. Many different types of substructure can be incorporated, such as the presence or absence of particular elements, rings and common functional groups. It is also possible to assign bits which encode how many occurrences of a particular feature exist. Structural keys used by the MACCS and Isis systems from Molecular Design are the best know of this type of bitstring.

### **5.2.2 Hashing Fingerprint**

Hashing fingerprint is a second commonly used type of binary screening, and does not require a predefined fragment dictionary; it uses an algorithmic approach to derive the bitstring. The Hashing fingerprint method produces all possible linear paths of connected atoms through the molecule containing between 1 and a pre-defined number of atoms. Each path defines a pattern of atoms and bonds which serves as the input to a pseudo-random number generator, which produces a set of bits which are then set to the value 1. The hashing process typically sets 4 or 5 bits per pattern. A bitstring might contain 1024 bits and after all paths have been examined a typical organic, drug-like molecule might have a total of 200-300 bits set to 1. Hashed fingerprints are used in a number of database systems and are particularly associated with the systems from Daylight Chemical Information Systems.

When using a bitstring screen, one first calculates the corresponding bitstring for the substructure query. Next, the query bitstring is compared with the bitstrings for all the molecules in the database. A molecule can only possibly match the query if it contains a 1 for every position in the bitstring where the query also has a 1. Well-designed screens can eliminate up to 99% of the molecular during this phase. After eliminating molecules that could not match the query, an atom-by-atom search for the molecules in conducted. The algorithm represents the molecular

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

graphs of both the query substructure and the potential molecular match by an adjacency matrix, which is a square, symmetric matrix such that the element  $(ij)$  has the value 1 if atoms  $i$  and  $j$  are bonded, and zero otherwise. The Ullmann algorithm tries to find matrices  $A$  such that  $A(AM)^T$  is identical to  $S$ , where  $M$  is the adjacency matrix of the molecule and  $S$  is the adjacency matrix of the substructure.

### **5.2.3 Database Searching – Conformational Properties and Functionality Features**

A 3D database search allows one to identify molecules that satisfy the chemical and geometric requirements of the receptor. A 3D database contains information about the conformational properties and functionality features of the molecules contained within it. There are two general types of 3D database searches. The choice of which to use depends on the information available about the target receptor. Pharmacophore mapping is used when an experimental structure of the target macromolecule is not available. Once a pharmacophore has been developed, it can then be used to find or suggest other active molecules. A pharmacophore refers to a set of features that is common to a series of active molecules. Such features are referred to as pharmacophoric groups, functional groups or molecules with similar physical and chemical properties such that they produce generally similar biological properties [Thornber 1979; Patani and LaVoie 1996]. A 3D pharmacophore specifies the spatial relationship between the groups. These relationships are often expressed as distances or distance ranges but may also include other geometric measures such as angles and planes.

There are two problems to consider when calculating 3D pharmacophores. First, unless the molecules are all completely rigid, one must take account of their conformational properties. Second, to determine which combinations of pharmacophoric groups are common to the molecules and can be positioned in a similar orientation in space. More than one pharmacophore may be possible. Some algorithms can generate hundreds of possible pharmacophores, which must then be evaluated to determine which best fits the data.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)

### 5.2.4 Constrained Systematic Search

Constrained systematic search address the problem of determining conformations in which the inhibitors can position multiple pharmacophoric groups in the same relative position in space. The constrained systematic search method of Dammkoehler, Motic and Marshall<sup>43</sup> showed that it is possible to determine what torsion angles of the rotatable bonds will enable conformations consistent with the previous results to be obtained.

A constrained system is represented by a matrix  $A$ . If  $A$  is untagged (has no input tags) and has  $r$  distinct output labels then  $A$  is a set of adjacency matrices with the form

$$A = [A_1 A_2 A_3 \dots A_r]$$

where  $A_i$  is an adjacency matrix that includes only edges with label  $i$ . That is,  $(A_i)_{x,y}$  is the number of edges from vertex  $y$  to vertex  $x$  with label  $i$ .

If  $A$  is tagged, with  $s$  distinct input tags, then  $A$  is a set of adjacency matrices with the form

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & \Lambda & A_{1,r} \\ A_{2,1} & A_{2,2} & \Lambda & A_{2,r} \\ \mathbf{M} & \mathbf{M} & & \mathbf{M} \\ A_{s,1} & A_{s,2} & \Lambda & A_{s,r} \end{bmatrix}$$

where  $A_{i,j}$  is an adjacency matrix that includes only edges with label  $i$  and tag  $j$ . That is,  $(A_{i,j})_{x,y}$  is the number of edges from vertex  $y$  to vertex  $x$  with label  $i$  and tag  $j$ .

### 5.2.5 Ensemble Distance Geometry

---

<sup>43</sup> Dammkoehler, R.A. Karasek, S.F., Shands, E.F.B., and Marshall, G.R. 1989. Constrained Search of Conformational Hyperspace. *Journal of Chemical Information and Computer Science* 32:244-255.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

Ensemble distance geometry can be used to simultaneously derive a set of conformations with a previously defined set of pharmacophoric groups overlaid. Ensemble distance geometry uses the same steps as standard distance geometry with the special feature that the conformation spaces of all the molecules are considered simultaneously<sup>44</sup>.

### **5.2.6 Clique Detection Methods**

It may be difficult to identify all possible combinations of the functional groups when many pharmacophoric groups are present in the molecule. Clique detection algorithms can be applied to a set of precalculated conformations of the molecules. Cliques are based upon the graph-theoretical approach to molecular structure. A clique is defined as a maximal completely connected subgraph. Finding the cliques in a graph is NP-complete. Many algorithms have been devised for finding cliques, including the method of Bron and Kerbosch<sup>45</sup>. The algorithm can be described as:

- Generate a family of low-energy conformations for the molecules
- Use the molecule with the smallest number of conformations as the starting point
- Use each of its conformations in turn and the reference structure
- Compare each conformation of every other molecule with the reference conformations and the cliques identified
- Obtain the cliques for each molecule by combining the results for each of its conformations
- Combine those cliques that are common to at least one conformation from each molecule to give a possible 3D pharmacophore for the entire set

### **5.2.7 Maximum Likelihood Method**

---

<sup>44</sup> Sheridan, R. P, Milakanton, R., Dixon, J.S and Venkataraghavan. 1986. The Ensemble Approach to Distance Geometry: Application to Nicotinic Pharmacophore. *Journal of Medicinal Chemistry* 29:899-906.

<sup>45</sup> Bron, C. and Kerbosch, J. 1973. Finding all Cliques of an Undirected Graph. *Communications of the ACM* 16:575-577.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

The maximum likelihood method eliminates the need for a reference conformation, effectively enabling every confirmation of every molecule to act as the reference. The algorithm scales linearly with the number of conformations per molecule, thus enable a large number of conformations to be handled.<sup>46</sup> The algorithm can be described as follows:

- Generate a set of conformations for each molecule
- Consider all possible combinations of pharmacophore features exhaustively
- Identify possible geometric arrangements of the features ins 3D space
- Score and rank according to how well the configuration describes the set of active molecules

For an implementation of the maximum likelihood method, see

<http://cmgm.stanford.edu/phylip/contml.html> for a description of CONTML, a program for estimating phylogenies by the restricted maximum likelihood method based on the Brownian motion model.

### **5.3 Molecular Docking**

Molecular docking attempts to predict the structure of the intermolecular complex formed between two or more molecules. Most docking algorithms are able to generate a large number of possible structures, and so they also require a means to score each structure to identify those of most interest. The docking problem involves many degrees of freedom. There are six degrees of translational and rotational freedom of one molecule relative to the other as well as the conformational degrees of freedom of each molecule.

Various algorithms have been developed to tackle the docking problem and can be characterized by the number of degrees of freedom they ignore. The simplest algorithms treat the two

---

<sup>46</sup> Barnum, D., Greene, J., Smellie, A. and Sprague, P. 1996. Identification of Common Functional Configuration among Molecules. *Journal of Chemical Information and Computer Science* 36:563-571.

**Federation of American Scientists**  
**Biomedical Computing Requirements for HPCS**  
**Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

molecules as rigid bodies and explore only the six degrees of translational and rotational freedom.

To perform conformationally flexible docking the conformational degrees of freedom need to be taken into account. All of the common methods for searching conformational space have been incorporated at some stage into a docking algorithm. Monte Carlo methods have been used to perform molecular docking, often in conjunction with simulated annealing.<sup>47</sup> Genetic algorithms can also be used to perform docking<sup>48</sup>, as well as distance geometry. An approach that is used by a number of programs involves the incremental construction of the ligand<sup>49</sup>. A typical incremental construction algorithm first identifies one or more base fragments within the ligand. The base fragments are docked into the binding site and may then be clustered to remove similar orientations. Each docked orientation of the base fragment(s) then represents the starting point for the conformational analysis of the rest of the ligand.

The ideal docking methods would allow both ligand and receptor to explore their conformational degrees of freedom. Molecular dynamics simulation of the ligand-receptor complex is one way to do this. However such calculations are computationally very demanding and typically used for refining structures produced using other docking methods.

Most docking algorithms generate a large number of potential solutions. Some of these can be rejected immediately because they have a high-energy clash with the protein. The rest are assessed using some scoring function. Many of the scoring functions attempt to approximate the binding free energy for the ligand binding to the receptor. Molecular mechanics is also widely used to calculate the energy of interaction. The calculation can be speeded up by pre-calculating electrostatic and van der Waals potentials on a regular grid that covers the binding site. The computational effort required to calculate the energy of interaction between ligand and protein is then linear in the number of atoms in the ligand, rather than being proportional to the product of

---

<sup>47</sup> Goodsell, D.S. and Loson, A.J. 1990. Automated Docking of Substrates to Proteins by Simulated Annealing. *Proteins: Structure, Function and Genetics* 8:195-202.

<sup>48</sup> Jones, G., Willett, P., Glen, R.C., Leach R. and Taylor R., 1997. Development and Validation of a Genetic Algorithm for Flexible Docking. *Journal of Molecular Biology* 267:727-748.

<sup>49</sup> Rarey M, Krammer B., Lengauer and Klebe, G. 1996. A Fast Flexible Docking Methods Using an Incremental Construction Algorithm,. *Journal of Molecular Biology* 261: 470-489.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)

the number of ligand atoms multiplied by the number of protein atoms<sup>50</sup>. Combining the results from more than one scoring function has been shown to give better results than using individual scoring functions on their own, an approach referred to as consensus scoring<sup>51</sup>

### 5.3.1 Protein-Ligand Docking in Drug Design

The first step of the drug design is to identify the lead structure, a small molecule which binds to a given target protein. The docking problems can be categorized as: Given two molecules with detailed 3-D structures:

- Do the molecules bind to each other?
- How strong is the binding?
- What does the binding complex look like?
- Protein-Protein or Protein-DNA docking: rigid-body docking, i.e., fixed overall shapes.
- Protein-Ligand docking: the ligand is not fixed in its overall shape
- Steric hinderance- does this complex exclude any other protein interaction?

Since most drugs are small molecules, protein-ligand docking is of great interest in pharmaceutical. The basic docking idea is to represent the active site by a set of spheres and then perform sphere matching<sup>52</sup>. There are two main algorithms, which are described below:

Algorithm 1: SPHGEN

- calculate the molecular surface
- generate spheres covering the active site
- cluster spheres, remove very similar ones
- radius too large

---

<sup>50</sup> Meng E.C, Shoichet, B.K. and Kuntz, I.D. 1992. Automated Docking with Grid-Based Energy Evaluation. *Journal of Computational Chemistry*. 13:505-524.

<sup>51</sup> Chaifson, P.S., Corkery, J.J., Murcko, M.A. and Walter, W.P. 1999. Consensus Scoring. A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *Journal of Medicinal Chemistry* 42:5100-5109.

<sup>52</sup> Kuntz et al., *Journal of Molecular Biology*. Vol. 161, pp. 269.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

- select clusters defining the active site
- color spheres by properties

Algorithm 2: MATCH (calculate a matching between ligand atoms L and protein spheres K)

- two matches  $(l_1, k_1), (l_2, k_2)$  are *distance-compatible* if  $|d(l_1, l_2) - d(k_1, k_2)| \leq \epsilon$
- search for matchings  $M = \{(l_i, k_j)\}$  with  $\max_{i,j} |d(l_i, l_j) - d(k_i, k_j)| \leq \epsilon$
- Matching-Graph: nodes  $L \times K$ , edges between distance compatible nodes
- Matchings are cliques in the matching graph (cliques = completely connected subgraphs)

### **5.3.2 Rigid-Body Protein-Ligand Docking**

With rigid-body protein-ligand docking, the protein and the ligand are assumed to be rigid. The first and most widely used rigid-body protein-ligand docking algorithm is DOCK<sup>53</sup>. The algorithm can be described as follows:

- a set of spheres is created inside the active site,
- the sphere represents the volume which could be occupied by the ligand molecule
- the algorithm searches for ligands (represented by spheres) that match the spheres describing the active site

Docking methods use receptor-ligand interactions to suggest binding modes. This is accomplished by identifying regions of binding site liable to interact in a given way *e.g.* hydrophobic regions or H-bonds. These interactions are clearly important, but other factors also affect binding. *Scoring functions* attempt to use all such factors to rank docked complexes in order of tightness of binding. Different scoring functions vary in which terms they treat and exact form of treatment.

---

<sup>53</sup> Kuntz, *et al.*, *J. Mol. Biol.*, **161**, 269, 1982

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

DOCK comes with a very simple scoring function to complement simple shape-based docking algorithm. DOCK ignores solvation, conformation and entropic effects completely. It uses molecular mechanics method to estimate binding free energy – equivalent to binding enthalpy in this case. DOCK uses AMBER force field for binding electrostatics & sterics *i.e.*

$$\Delta G_{\text{bind}} \approx \Delta E_{\text{elec}} + \Delta E_{\text{vdw/ster}}$$

$\Delta E_{\text{vdw/ster}}$  includes attractive van der Waals interactions and repulsive steric clashes. These are calculated from the standard Lennard-Jones 6-12 potential using pairwise atom-atom terms.

The electrostatic term in DOCK is taken as a simple sum of charge-charge interactions. Charges are estimated by Gasteiger's electronegativity equalisation scheme – fast route to charges from 2D structure only. Dielectric shielding can be applied to above equation to model the shielding of charges by each other. The value of  $\epsilon$  varies for different receptors.

The simplicity of DOCK is attractive – it is very quick to evaluate & easy to interpret. Calculation can be speeded up by evaluating electrostatic & conformation 'fields' on a grid within the binding site. The score for a given ligand is then easy to calculate from atomic positions in the grid. The lack of entropic and conformation effects means DOCK is only applicable to series of similar ligands. In spite of this limitation, DOCK is remarkably successful.

A different approach to empirical scoring is the 'Potential of Mean Force' (PMF) function of Muegge & Martin. Atom types for important interactions in complexes are identified. The strength of atom-atom interactions is found by regression against binding energies including distance-dependence terms. This includes solvation, entropic *etc.* effects implicitly, and therefore is very fast.

Recent studies suggest that no single scoring function works for every problem. Two main measures of quality are used:

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

- comparison with ranking from experimental binding energies
- agreement with X-ray structure (RMSD)

Certain interactions are better represented by different scoring functions. Simple DOCK approaches can work better than more complex ones. This leads to *consensus scoring*, where several scoring functions are used at once. The simplest approach is to take the average predicted binding energies – but this doesn't work well as results are not always on same scale. Rank order (*i.e.* 1st,2nd,3rd,...) has been shown to be better than predicted DGbind. Several statistics for consensus scoring have been proposed from rank. The first study used DOCK, GOLD, FlexX and PMF to rank 15 ligands from 1st-15<sup>th</sup>. Best criteria found to be 'worst-best rank' and 'rank-sum'. Worst-best drops the worst of the 4 ranks and takes next worst, while rank-sum adds the remaining 3 ranks *e.g.* ligand has ranks 3, 5, 6 & 12 – worst best is 6th and rank-sum is 14. This method is known as CScore – and is implemented in several packages now.

### **5.3.3 Flexible ligand docking**

GOLD, a genetic optimization for ligand docking is a program developed by Gareth Jones at the University of Sheffield (Sheffield, UK) in collaboration with Glaxo Wellcome (London) and the Cambridge Crystallographic Data Centre (CCDC; Cambridge, UK) where the technique is applied to the problem of docking ligands to protein binding sites<sup>54</sup>. A *chromosome* describing the conformation of the ligand and selected protein side chains by defining the torsion angle of each rotatable bond. Another *chromosome* stores a mapping between hydrogen bond partners in the protein and the ligand. 3D structures are generated from these two chromosomes. A scoring function that evaluates the hydrogen bond, ligand internal energy and van der Waals energy is applied as the fitness function. The GOLD docking method also has its own scoring function, which is slightly more sophisticated than DOCK. Rather than a simple electrostatic term, GOLD models the H-bonds in a complex. Careful studies of how small molecules interact and

---

<sup>54</sup> Jones et al., *Journal of Molecular Biology* 267:727-748, 1997.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

crystallise give geometry/energy rules for different H-bonds. GOLD is slower to evaluate than DOCK, but its better search capability results in comparable performance.

FlexX is perhaps the most complex scoring function currently available. All the scoring functions discussed so far attempt to calculate binding energy directly. All employ some version of the ‘master equation’ which partitions DGbind. An alternative approach is to find an ‘empirical scoring function’ from a database of binding energies. Empirical scoring functions are a form of QSAR, with DGbind in place of activity. These calculate properties of ligands and train the QSAR/scoring function from known values using linear regression and/or PLS. Properties used include polar/non-polar SASA, H-bond donors and acceptors and number of rotatable bonds.

### **5.3.4 Docking by simulation**

Molecular dynamics simulations use the force field to calculate the forces on each atom of the whole system. Following classical mechanics, velocities and accelerations are calculated and the atoms are moved slightly with respect to a given time step. Simulated annealing is another optimization algorithm that avoids getting into local minima but lacks physical interpretation of the simulation itself.

### **5.4 De novo drug design**

Docking/scoring finds active molecules from a list of possible ligands. *De novo* drug design attempts to find new structures rather than comparing new ones. While database searching is an attractive way to discover new lead compounds, database searching does not provide molecules that are structurally novel. In addition, many databases are biased towards particular classes of compounds, and so limit the range of structures that can be found. In de novo design, the 3-D structure of the receptor or the 3D pharmacophore is used to design new molecules. The starting point is a receptor site from X-ray or modelling. However, instead of possible ligand molecules, a database of common & realistic fragments is searched for a fit with binding site. Fragments are chosen to give good shape and interaction overlap with binding site. Scoring functions can be

**Federation of American Scientists**  
**Biomedical Computing Requirements for HPCS**  
**Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

used to define which fragment(s) are best suited. Fragments are then re-combined somehow to give possible drug leads. There are two basic types of do novo design algorithm: ‘outside-in’ or ‘inside-out’. Both look to grow a fragment within the binding site, but differ in how fragments are chosen and molecule re-combined. These will ideally converge to the same (or similar) solution given the same database and scoring function. This is a relatively new approach, so not yet clear if either has any advantages.

### **5.4.1 Outside-In Method**

The outside-in method<sup>55</sup> finds fragments which bind tightly to regions of active site, which are then combined to real molecules. Initially the binding site is sampled for ‘*site points*’ where interactions could occur. Next, fragments are placed on site points and scored with some function. The scoring function and systematic search can usually identify binding fragments. This is only half the problem – combining fragments is not trivial. One solution is to use a database of common connectors, and match geometry of fragments to known linker groups. Alternatively a ‘skeleton’ can be grown between fragments using rings and/or acyclic C—C bonds. For a discussion of the algorithm see <http://www.andrew.cmu.edu/user/sowen/abstracts/Wa855.html>

### **5.4.2 Inside-Out Method**

The inside-out method takes the opposite approach. First one starts with a central ‘scaffold’ fragment and incrementally grows fragments off this. The scaffold fragment should be rigid and tightly bound, but this is not always obvious. Each time a fragment is added, the resulting molecule is re-docked into binding site. Once docked the new molecule’s binding energy can be predicted using one or more scoring functions. The search stops either when no further improvement found, or when binding is tighter than some pre-defined cutoff. Repeated docking and scoring can be slow, but can discover new ligands.

---

<sup>55</sup> Lewis, R.M. and Leach A. R, 1994. Current Methods for Site-Directed Structure Generation. Journal of Computer-Aided Molecular Design \*:467-475.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

The main goal of rational or *de novo* design is to find an active compounds for further development, or *leads*. Lead optimisation is equally important in overall drug development process. Enhancement of activity is a major goal, but better solubility, delivery/distribution, toxicity & synthesis also important. This often takes the form of ‘virtual screening’ of possible targets before any synthesis.

### **5.5 Software for automated docking**

A number of popular software packages for drug design are available and used widely, including:

**GOLD:** <http://www.ccdc.cam.ac.uk/prods/gold.html>

**AutoDock:** <http://www.scripps.edu/pub/olson-web/doc/autodock/>

**DOCK:** <http://www.cmpharm.ucsf.edu/kuntz/dock.html>

**DockVision:** <http://www.dockvision.com>

**FlexX:** <http://cartan.gmd.de/FlexX>

**ICM:** <http://www.molsoft.com/products/modules/dock.htm>

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)

## 6 Systems Biology

Molecular biology is mainly focused on identification of genes and functions of their products, which are components of the system. The next major challenge is to understand at the system level biological systems that are composed of components revealed by molecular biology<sup>56</sup>. The goal is to understand biological systems within a consistent framework of knowledge built up from the molecular level to the functional system level. Understanding biology at the system level – not only gene networks, but also protein networks, signaling networks, metabolic networks and specific systems such as the immune system or neuronal networks will be a major driver of HPCS requirements in the coming years.

At a very abstract level, a cell can be divided into two general subnetworks, a regulatory network and a metabolic network (Kremling, et al., 2000). These networks possess very different characteristics. The metabolic network is mainly occupied with substance transformation to provide metabolites and cellular structures. The regulatory network's main task is information processing for the adjustment of enzyme concentrations to the requirements of variable internal and external conditions. This network involves the use of genetic information.

<b>Cell Network</b>	<b>Task</b>	<b>Examples</b>
Metabolic pathway	Enzyme reactions on chemical substances	Intermediary metabolism, secondary metabolism, macromolecular metabolism
Regulatory pathway	Macromolecular reactions and interactions Direct protein-protein interactions and gene expressions	Membrane transport, signal transduction, ligand-receptor interaction, cell cycle, cell death

<sup>56</sup> Hiroaki Kitano, Foundations of Systems Biology, p. 1.

**Federation of American Scientists**  
**Biomedical Computing Requirements for HPCS**  
**Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

The complex network of biochemical reaction/transport processes and their biochemical spatial organization make the development of a predictive model of living cell a “grand challenge” problem. Cell signaling, cell motility, organelle transport, gene transcription and translation, morphogenesis and cellular differentiation cannot easily be accommodated into existing computational frameworks. Conventional approaches using the numerical integration of continuous, deterministic rate equations can provide useful when systems are large or when molecular details are of little importance. However when the resolution of experimental techniques increases, conventional models become unwieldy. Difficulties include the importance of spatial location within the cell, the instability associated with reactions between small numbers of molecular species and the combinatorial explosion of large numbers of different species. For example, signaling pathways commonly operate close to points of instability and frequently employ feedback and oscillatory reaction networks that are sensitive to the operation of small numbers of molecules. Gene transcription is controlled by small assemblies of proteins operating in an all-or-none fashion and post-translational modifications, so that whether a specific protein is expressed or not is to some extent a matter of chance (Ko, 1991; Kingston and Green, 1994; Tijan and Maniatis, 1994; McAdams and Arkin, 1999). Stochastic methods are being used. The idea is to represent individual molecules rather than the concentrations of molecular species and to apply Monte Carlo methods to predict their interactions. In the stochastic modeling approach, rate equations are replaced by individual reaction probabilities and the output has a physically realistic stochastic nature. Techniques are available by which large numbers of related species can be coded in an economical fashion and key concepts such as signaling complexes and heat-driven flipping of protein conformations can be embodied in the program. (Shimizu and Bray, p. 215).

In the cell, various components interact in diverse manners. All cellular subsystems are highly nonlinear, and subsystem couplings are often nonlinear as well. This nonlinearity indicates that the whole system is not equivalent to the sum of its subsystems. A cell must be treated as a whole in order to elucidate a cell’s real behavior and role it a part of the whole. Cell simulators

**Federation of American Scientists**  
**Biomedical Computing Requirements for HPCS**  
**Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

must therefore allow simulation of cell subsystems in both isolated and coupled forms. To simulate coupled subsystems, it is necessary to perform computations on mutually interacting subsystems with different computational properties on a single platform. There is, however, no universal algorithm that can efficiently simulate all subsystems at once, so simulators must allow multiple computation algorithms to coexist in a single model. To support mixed-mode computation, the software must therefore provide a single abstract programming interface that both allows indiscriminate interaction among modules and gives front-end programs a standard means of visualizing and manipulating these modules. Implementations of the algorithm modules must also be isolated from the system-provided common interface. Understanding biological systems requires:

- identification of the structures of the system – primarily regulatory relationships of genes and interactions of protein that provide signal transduction and metabolism pathways, as well as the physical structure of organisms, cell, organelle, chromatin and other components. Both the topological relationship of the network of components as well as parameters for each relation needs to be identified. Identification of gene regulatory networks for multicellular organisms is even more complex as it involves extensive cell-cell communication and physical configuration in 3-D space.
- analysis of system behavior – once a system structure is identified, its behavior needs to be understood
- a method to control the state of biological systems
- design of biological systems with the aim of providing cures for diseases
- sort through redundancies in the system

Simulations need to be able to simulate gene expression, metabolism and signal transduction for a single and multiple cells. The simulations must be able to simulate both high concentrations of proteins that can be described by differential equations and low concentrations of proteins that need to be handled by stochastic process simulation. Some efforts on simulating a stochastic process (McAdams and Arkin, 1998) and integrating it with high concentration level simulation are underway. In some cases the model requires not only gene regulatory networks and

**Federation of American Scientists**  
**Biomedical Computing Requirements for HPCS**  
**Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

metabolic networks, but also high-level structures of chromosomes such as heterochromatin structures.

The simulations need to be coupled with parameter optimization tools, a hypothesis generator and a group of analysis tools. The algorithms need to be designed precisely for biological research. For example, the parameter optimizer needs to find as many local and global minima as possible because there are multiple possible solutions of which only one is actually used. The assumption that the most optimal solution is used in an actual system does not hold true in biological systems. The tools and analysis required are:

- A database for storing experimental data
- A cell and tissue simulator
- Parameter optimization software
- Bifurcation and systems analysis software
- Hypotheses generator and experiment planning advisor software, and
- Data visualization software

### **6.1 Systems Biology Analysis Methods**

Modern metabolic pathway models typically consist of primary state variables for molecular species concentrations, one ODE for each enzyme reaction, and a stoichiometric matrix.

Researchers derive the rate equations of most modern enzyme kinetics models using the King-Altman method.<sup>57</sup> Additional algebraic equations commonly serve as constraints on the system. Thus, most metabolism models are described as differential-algebraic equations (DAEs). factors, polymerases, and genes. These lowcopy- number molecules organize gene expression in a highly stochastic fashion. For example, the stochastic behavior of gene expression's initial phase has binary consequences: binding a rare transcription factor and a single gene in a cell can determine whether the gene is turned on or off. Stochastic modeling is an approach to modeling

---

<sup>57</sup> E.L. King and C. Altman, "A Schematic Method of Deriving the Rate Laws for Enzyme Catalysed Reactions," *J. Physical Chemistry*, vol. 60, 1956, pp. 1375-1378.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)

phenomena such as intracellular signaling and gene expression. The conventional approach of representing biochemical reactions by continuous, deterministic rate equations cannot be easily applied to intracellular processes based on multiprotein complexes or those that depend on the individual behavior of small numbers of molecules.<sup>58</sup>

StochSim is a stochastic biochemical simulator that represents individual molecules and molecular complexes as individual software objects. It employs a unique algorithm that effectively simulates chemical systems such as the bacterial chemotaxis-signaling pathway, where only a few molecules are involved and some of them are multistate. MCell, another simulator, treats molecules individually rather than statistically, with a Monte Carlo-type random-walk algorithm for Brownian dynamics. MCell simulates interactions between ligands and binding sites on receptors, enzymes, and transporters (among other molecules). The high computational costs of both StochSim and MCell discourage their use for simulation of subcellular dynamics, yet their algorithms are likely to become indispensable if given appropriate roles in a mixed-mode whole-cell model. The following table summarizes cellular processes and typical computational approaches for each:

**Cellular processes and typical computational approaches<sup>59</sup>**

Process type	Dominant phenomena	Typical computation schemes
Metabolism	Enzymatic reaction	DAE, S-Systems, FBA
Signal transduction	Molecular binding	DAE, stochastic algorithms (StochSim and Gillespie, for example), diffusion-reaction
Gene expression	Molecular binding, polymerization, degradation	OOM, S-Systems, DAE, Boolean networks, stochastic algorithms
DNA replication	Molecular binding, polymerization	OOM, DAE
Cytoskeletal	Polymerization, depolymerization	DAE, particle dynamics
Cytoplasmic streaming	Streaming	Finite-element method
Membrane transport	Osmotic pressure, membrane potential	DAE, electrophysiology

DAE—differential-algebraic equations (rate equation-based systems), FBA—flux balance analysis, and

<sup>58</sup> Shimizu, Thomas Simon and Dennis Bray, *Computational Cell Biology – The Stochastic Approach*, p. 213.

<sup>59</sup> Kouichi Takahashi, Katsuyuki Yugi, Kenta Hashimoto, Yohei Yamada, Christopher J.F. Pickett, and Masaru Tomita, *Computational Challenges in Cell Simulation: A Software Engineering Approach*. <http://ecell.sourceforge.net/ieeetak.pdf>

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

OOM—object-oriented modeling (includes E-Cell's substance-reactor model, or SRM)

Modeling of macromolecular interactions permits simulation of signal transduction into, across, and out of cells. With multicellular models, it is possible to investigate chemical and mechanical processes that occur in tissues, ranging from the relatively simple lipid bilayers that generally modulate intracellular chemistry to the much more complex assemblages that make up such tissues. Commonly used analysis methods for systems biology are bifurcation analysis, metabolic control analysis and sensitivity analysis.

### **6.1.1 Simulation Software and Tools**

- Gepasi, is a rate-equation-based simulator and is widely used by biochemists for both research a
- Copasi, Gepasi's successor, focuses on large scalability and distributed parallel computing<sup>60</sup>.
- DBSolve is an ordinary differential equation- (ODE-) based simulator<sup>61</sup>.
- Promot is an object-oriented modeling environment that uses the Diva numerics solver as a simulation back end.<sup>62</sup>
- Scop (Simulation Control Program is a commercial tool for block-oriented generic simulation of complex systems, to run differential- and difference-equation-based cell models.<sup>63</sup>

---

<sup>60</sup> P. Mendes, "Gepasi: A Software Package for Modeling the Dynamics, Steady States, and Control of Biochemical and Other Systems," *Computer Applications in Biosciences (CABIOS)*, vol. 9, no. 5, Oct. 1993, pp. 563–571.

<sup>61</sup> I. Goryanin, T.C. Hodgman, and E. Selkov, "Mathematical Simulation and Analysis of Cellular Metabolism and Regulation," *Bioinformatics*, vol. 15, no. 9, Sept. 1999, pp. 749–758.

<sup>62</sup> M. Ginkel et al., "Application of the Process Modeling Tool Promot to the Modeling of Metabolic Networks," *Proc. 3rd MathMod*, I. Troch and F. Breitenecker, eds., vol. 2, ARGESIM report no. 15, vol. 2, Vienna Univ. of Technology, Vienna, 2000, pp. 525–528.

<sup>63</sup> J.M. Kootsey et al., "Scop: An Interactive Simulation Control Program for Micro- and Minicomputers," *Bull. Math. Biology*, vol. 48, nos. 3–4, 1986, pp. 427–441.

**Federation of American Scientists**  
**Biomedical Computing Requirements for HPCS**  
**Draft Preliminary Report**

Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)

- A-Cell is a GUI-based software for constructing biochemical reactions and electrophysiological models of neurons and other cell types.<sup>64</sup>
- Bio-Spice, initially intended for genetic circuit simulation is now being developed as a generic modeling and simulation environment linked to object-relational databases.<sup>65</sup>
- Jarnac, or Scamp II, provides a flexible scripting language that supports dynamic object-oriented cell models.<sup>66</sup>
- Virtual Cell lets modelers construct spatial and biochemical models in both biological and mathematical semantic planes, and has support for empirical 3D data from microscopy.<sup>67</sup>
- CellML is a modeling language similar to Virtual Cell that is being developing by the University of Auckland and Physiome Sciences sre developing a similar modeling language called CellML ([www.cellml.org](http://www.cellml.org)).
- StochSim is a stochastic biochemical simulator that represents individual molecules and molecular complexes as individual software objects.<sup>68</sup>
- MCell, also a simulator, treats molecules individually rather than statistically, with a Monte Carlo-type random-walk algorithm for Brownian dynamics.<sup>69</sup>

---

<sup>64</sup> K. Ichikawa, "A-Cell: Graphical User Interface for the Construction of Biochemical Reaction Models," *Bioinformatics*, vol. 17, no. 5, May 2001, pp. 483–484.

<sup>65</sup> H.H. McAdams and A. Arkin, "Simulation of Prokaryotic Genetic Circuits," *Ann. Rev. of Biophysics and Biomolecular Structure*, vol. 27, 1998, pp. 199–224.

<sup>66</sup> H.M. Sauro, "Scamp: A General-Purpose Simulator and Metabolic Control Analysis Program," *Computer Applications in Bioscience*, vol. 9, no. 4, Aug. 1993, pp. 441–450.

<sup>67</sup> L.M. Loew and J.C. Schaff, "The Virtual Cell: A Software Environment for Computational Cell Biology," *Trends in Biotechnology*, vol. 19, no. 10, Oct. 2001, pp. 401–406.

<sup>68</sup> N. Le Novere and T.S. Shimizu, "Stochsim: Modeling of Stochastic Biomolecular Processes," *Bioinformatics*, vol. 6, no. 6, June 2001, pp. 575–576.

<sup>69</sup> T.M. Bartol, Jr., et al., "MCell: Generalized Monte Carlo Computer Simulation of Synaptic Transmission and Chemical Signaling," *Society*

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

### **6.1.1 Integrative Software Environments**

The ERATO Systems Biology Workbench project is to create an integrated software environment that permits sharing of models and resources between simulation and analysis tools for systems biology. The initial focus is on achieving interoperability between seven leading simulation tools: BioSpice (Arkin, 2001), DBSolve (Goryanin, 2001), E-Cell, Gepasi (Mendes, 1997,2001), Jarnac (Sauro, 1991; Sauro and Fell, 2000), StochSim (Bray et al., 2001; Morton-Firth and Bray, 1998), and Virtual Cell (Schaff, et al., 2000, 2001). As part of the effort, the project has also developed a model description language, the Systems Biology Markup Language (SBML) that can be used to represent models in a form independent of any specific simulation/analysis tool. SBML is a versatile and common standard that enables the exchange of data and modeling information among a wide variety of software systems (Hucka et al., 2000, 2001). It is an extension of XML, and is expected to become the industrial and academic standard of the data and model exchange format.

The Virtual Laboratory uses a process modeling tool PROMOT, originally designed for application in chemical engineering. It allows for the computer-aided development and implementation of mathematical models for living systems (Ginkel et.al., 2000). For the numerical analysis of the resulting models the simulation environment DIVA (Mangold et al., 2000) is used. DIVA deals not only with large-scale differential-algebraic systems which arise in chemical process engineering but also in the mathematical modeling of complex cellular networks. Inside DIVA many different numerical computations can be performed based on the same model, including dynamic and steady state simulation, parameter estimation, optimization and the analysis of nonlinear dynamics. There are currently four methods of special interest for cellular models:

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

- Dynamic simulation of the models with different integration algorithms
- Sensitivity analysis for parameters with respect to experimental data
- Parameter identification according to experimental data
- Model-based experimental design.

Most numerical algorithms in DIVA are taken from professional numerical libraries like HARWELL and NAG. The system also has additional methods like steady state continuation and bifurcation analysis. The visualization and postprocessing are done using MATLAB.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

## **7 Computer-aided diagnostic imaging and image-guided interventions**

Emerging new imaging technologies could very likely change the face of medicine, making it increasingly possible to noninvasively detect, diagnose, and guide therapy for a large variety of diseases. Computer-aided diagnostic imaging and image-guided interventions are used for monitoring of disease progression, diagnosis, preoperative planning and intraoperative guidance and monitoring. Imaging is being used to pinpoint signifying events that mark disease onset and define its biologic characteristics. Applications include use of multiple imaging modalities for diagnosis and position of the lesion in three-dimensional space, real-time imaging of anatomy, physiology (e.g., blood flow, neuron activation) and function during surgery, video imaging in laparoscopic surgery, image-guided placement of catheters and other devices, and surgical computer-aided design and distance medicine.

### **7.1 Applications**

Image processing is being used increasingly to aid in surgical planning to facilitate selection of the appropriate intervention, assess operative risk in surgical cases, allow the physician to visualize the normal and pathologic relationships as well as select the surgical approach non-invasively and preoperatively, and permit the surgeon to localize specific objects of interests, such as lesions, intraoperatively in conjunction with video registration. The models are also increasingly used preoperatively to enhance resident training.

Surgical guidance systems integrate capabilities for data analysis and on-line interventional guidance into the setting of interventional MRI. Various pre-operative scans, such as MRI, MR angiography, and functional MRI are fused and automatically aligned with the operating field of the interventional MR system. Both pre-surgical and intra-operative data are typically segmented to generate three-dimensional surface models of key anatomical and functional structures. The models are combined in a three-dimensional scene along with reformatted slices that are driven by a tracked surgical device. This allows the pre-operative data to augment interventional imaging to expedite tissue characterization and precise localization and targeting. As the surgery

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

progresses, and anatomical changes necessitate that the interventional data be refreshed for software navigation in true real time. The entire process, from scanned data to 3D model can require 3-15 hours per patient.<sup>70</sup> A pipeline approach is typically used. The Surgical Planning Lab at Brigham and Women's Hospital<sup>71</sup> uses a such an approach, which consists of the following processes: Each image is preprocessed to reduce noise using anisotropic diffusion filtering. Next, segmentation, based on signal intensities and a voxel connectivity is performed. Manual editing may be used to adjust the initial segmentation; new methods are being developed to more fully automate the process. The resulting labeled images, 3D objects of the related structures are reconstructed using the marching cubes algorithm and a surface rendering method. These objects are integrated, and displayed in 3-D using that software that allows each object to be individually colored, made translucent, removed, rotated, translated and scaled as the viewer wishes.

Recent advanced in imaging research have shown the potential to change many aspects of clinical medicine within the decade. New imaging techniques will enhance the ability to diagnose disease and to monitor drug delivery and effect; for instance, with new imaging techniques that will provide surrogate markers of both treated and untreated disease. Characterization of biological factors (biomarkers) of various conditions may offer useful measures that can predict the course of disease. As valid indicators of disease, biomarkers may serve as candidate surrogate endpoints in clinical trials of novel interventions. Developing and evaluating new drugs and medical therapies in less time and at lower cost is of enormous potential benefit for modern healthcare. Biomarkers are routinely visualized by current imaging modalities, with exciting research promising even more comprehensive evaluation in the near future. Beyond their impact on patient care, biomarkers may be used as preclinical and clinical endpoints in evaluating the safety and effectiveness of new drug, biologic and device-based medical therapies and speeding the regulatory evaluation of new treatments. Examples of

---

<sup>70</sup> <http://splweb.bwh.harvard.edu:8000/pages/papers/shin/ns/ns.html#Introduction>.

<sup>71</sup> <http://splweb.bwh.harvard.edu:8000/pages/aboutspl/about.html>

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

successful application of biomarker data to therapeutic evaluation include Betaseron for use against multiple sclerosis and Herceptin in the treatment of breast cancer<sup>72</sup>.

The ability to store and integrate imaging and medical information along with tools for management of imaging data will enable more accurate and timely diagnosis of disease and improved patient care. Radiologists will be able to search for the patient in the database and then retrieve all of the patient's images on the same computer. Digital images will allow for a wide variety of measurements to be performed, including uni-dimensional, bidimensional and volumetric. The increasing complexity of information available from image data sets increases demand on the diagnostic skills of radiologists. Drug trials present a significant opportunity for smarter imaging processes. Drug studies involve many hundreds of patients, and typically requires multiple scans of each patient. For example, an evaluation of a new drug may involve 1,000 patients, each of which are scanned three times, resulting in 3,000 sets of data. Two ways to improve diagnostic performance are by improving the radiologist's accuracy and by increasing the utility of diagnostic decisions. The ability to perform multimodal image fusion (eg, combine data sets from PET and CT or SPECT and MRI) increases complexity and also requires innovative methods for increasing diagnostic accuracy, such as feature analysis and computer-aided diagnostic tools. Statistical prediction rules are a form of computer-based decision support that improves diagnostic accuracy. Such rules can enable analysis of more than 20 variables on a mammogram and combine the results to provide an estimate of the probability of cancer. These tools are powerful and can improve the quality and accuracy of diagnostic techniques, as illustrated by application of MRI for staging prostate cancer.<sup>73</sup>

## **7.2 Techniques and Methods**

High performance computing is becoming increasingly important in imaging applications because radiological image analysis is undergoing a shift, from the visual analysis of collections

---

<sup>72</sup> Website biomarkers, <http://www.biomarkers.org/NewFiles/about.html>

<sup>73</sup> Seltzer SE, Getty DJ, Tempany CM, et al. Staging prostate cancer with MR imaging: a combined radiologist-computer system. *Radiology*. 1997;202:219-226.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

of planar images, to the computerized quantitative analysis of volumetric images<sup>74</sup>. Diagnostic medical images, acquired from a number of modalities, are now providing three-dimensional raster data. Another increasingly important, computationally demanding, area is that of intraoperative imaging and image guided surgery. The analysis of imaging data is increasingly limited by memory and speed constraints. In the development of automated systems and systems which operate during surgery, the speed at which one data set can be processed is an issue. Other clinical considerations, such as the time between a scan and surgery, place a limit on the time available for a single registration. Rapid registration enhances the design and development process for new image analysis systems, since during development changes and novel enhancements require rapid feedback for extensive testing to be practicable. Consequently, it is important to have a rapid, low latency, registration technique.

Key issues for digital imaging are segmentation and registration. Signal processing techniques are used to enhance features and generate the desired segmentation. Results of the segmentation are aligned to other data acquisitions and to the actual patient during procedures<sup>75</sup>. Results of the segmentation are visualized using different rendering methods. The following sections describe typical steps in the digital imaging pipeline, and the methods and techniques employed.

### **7.2.1 Feature Enhancement**

Image data is filtered prior to segmentation to reduce the noise level and to emphasize image structures of interest. Segmentation of MR images often uses anisotropic diffusion for enhancing the gray-level image structure prior to segmentation<sup>76</sup>. By smoothing along structures and not across, the noise level can be reduced without severely blurring the image. Steerable filters that conform to the local structure adaptively are often used<sup>77</sup>.

---

<sup>74</sup> <http://splweb.bwh.harvard.edu:8000/pages/papers/warfield/hpc-reg/hpcreg.html#rubin>

<sup>75</sup> Jolesz, F. A. 1997. Image-guided Procedures and the Operating Room of the Future. *Radiology* 204:601-612.

<sup>76</sup> Greig, G., O. Kuebler, R. Kikinis, and F.A. Jolesz. 1992. Nonlinear Anisotropic Filtering of MRI Data. *EIII Trans. Medical Imaging*. 11 (2):221-232.

<sup>77</sup> Granlund G. H. and H. Knutsson. 1995. *Signal Processing for Computer Vision*. Kluwer Academic Publishers.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

Convolution involves multiplication and summation of filter kernel coefficients with signal voxels, over the local area that the filter supports. Since the result in each voxel can be calculated independently, these calculations can be done in parallel and thus the speedup for convolution is linear with the number of CPUs. For large filter kernels (e.g., 9x9x9 voxels), it is more efficient to calculate the result of a convolution using the Discrete Fourier Transform (DFT).

For example, FFTW is a software package developed at MIT.<sup>78</sup> FFTW is a C subroutine library for performing the Discrete Fourier Transform (DFT) in one or more dimensions. An MPI version of the FFTW routines is available which makes it possible to perform the FFT calculation on distributed memory machines in addition to shared-memory architectures.

### **7.2.2 Classification**

Classification is a technique for the segmentation of medical images. The k-Nearest Neighbor (k-NN) classification rule is a technique for nonparametric supervised pattern classification. For a description see Duda, 1973<sup>79</sup> describes k-NN classification and its properties. Each voxel is labeled with a tissue class selected from a set of possible classes. The possible tissue classes are described, in k-NN classification, by selecting a set of typical voxels (prototypes) for each tissue type. Voxels of an unknown class are then classified by comparing the voxel intensity characteristics with those of the prototypes and selecting the class that occurs most frequently among the k nearest prototypes.

The classification of each voxel is independent of neighboring voxels. The most straightforward parallelization strategy is to apply the k-NN classification rule to several voxels at the same time, up to the number of CPUs available for computation. Speedup is linear with the number of CPUs.

---

<sup>78</sup> Frigo, Matteo and Steven G. Johnson. 1997. The Fastest Fourier Transform in the West. Proceedings of the 1998 International Conference on Acoustics, Speech and Signal Proceedings. 1998.

<sup>79</sup> Duda R.O. and P.E. Hart. 1973. Pattern Classification and Scene Analysis, John Wiley & Sons Inc.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)

### 7.2.3 EM Segmentation

EM segmentation is a method that iterates between conventional tissue classification and the estimation of intensity inhomogeneity to correct for imaging artifacts. The EM algorithm consists of a conventional classification step, an intensity prediction step, and an intensity correction step. Classification is parallelized by classifying different voxels simultaneously, as above. The same is done with the intensity prediction step. Intensity correction primarily involves low-pass filtering implemented with a parallel unity gain filtering step<sup>80</sup> that costs only two multiplies per voxel per axis, independent of filter length

### 7.2.4 Linear Registration

Registration is a universal problem which must be addressed in almost all computer-assisted or image-guided surgical systems. It is equally important in orthopaedics, neurosurgery, spine surgery, cranio-facial surgery, or any speciality in which computer-assisted surgical techniques are employed. Linear registration algorithms align several complementary data sets of the same subject (e.g., a CT and an MRI scan). Another application is the initial alignment, as a preliminary step before non-linear registration, of a canonical data set and the data from a specific subject. Different algorithms that have been published in the literature Warfield, et al 1998, 2002<sup>81 82</sup>, West, et al 1997<sup>83</sup>, typically trade off speed (e.g., through feature extraction or subsampling) and robustness and capture range (e.g., by simulated annealing).

---

<sup>80</sup> Wells WM, W.E. L. Grimson, R. Kikinis, and F.A. Jolesz. Adaptive Segmentation of MRI Data. IEEE Transactions on Medical Imaging. 15 (4):429-442.

<sup>81</sup> Warfield, S.K. / Jolesz, F.A. / Kikinis, R.,. A high performance computing approach to the registration of medical imaging data, *Parallel Computing*, Sep 1998.

<sup>82</sup> Warfield, S. Guimond, A. Roche, A., Bharatha, Tei, A., Talos, F., Rexilius, J. Ruiz-Alzola, J., Westin, J, Haker, S., Angenent, S., Tannenbaum, A, Jolesz, F., Kikinis, R. Brain Mapping: The Methods, Second Edition, as chapter 24 on pages 661–690, Academic Press of San Diego, USA in 2002.

<sup>83</sup> West J, Fitzpatrick JM, Wang MY, Dawant BM, Maurer CR, Kessler RM, Maciunas RJ, Barillot C, Lemoine D, Collignon A, Maes F, Suetens P, Vandermeulen D, van den Elsen PA, Napel S, Sumanaweera TS, Harkness B, Hemler PF, Hill DLG, Hawkes DJ, Studholme C, Maintz JBA, Viergever MA, Malandain G, Pennec X, Noz ME, Maguire GQ, Pollack M, Pelizzari CA, Robb RA, Hanson D, Woods RP *Comparison and Evaluation of Retrospective Intermodality Brain Image Registration Techniques* J. Comp. Assist. Tomogr. Vol 21, No 4, pp. 554-566. 1997.

**Federation of American Scientists**  
**Biomedical Computing Requirements for HPCS**  
**Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

The registration of different images of the same patient is referred to as inpatient registration. A number of registration algorithms have been successfully applied to this problem. Inpatient registration is used for the integration of scans of a patient from multiple imaging modalities (such as PET, SPECT, CT and MRI) and for the analysis of series of scans of a patient acquired over time. Interpatient registration involves the alignment of scans of different patients. Primary application has been the study of anatomical variation, to compare scans from different patients in a common coordinate system. Comparing and contrasting anatomical variability of normal volunteers and groups of patients can help to reveal the physical changes associated with certain diseases. Interpatient registration techniques can usually be used for inpatient registration. However, inpatient registration schemes often do not generalize successfully to interpatient registration.

Precise alignment of corresponding regions is possible when the same patient is being examined because the same regions appear and have the same shape and location. For inpatient registration, a rigid body transform is usually sought -- the alignment transform is constrained to consist of only translation and rotation parameters. For interpatient registration, this is usually not sufficient. Even though the scans to be aligned have the same anatomic structures, the shape and the relative location of these structures can be different.

For interpatient registration applications the different size of different subjects leads to the desire to also solve for scale parameters. A nine parameter transform (three translation, three rotation, and three scale parameters) is usually computed, allowing the capture of global scale, rotation and translation differences. After the application of such a nine parameter transform, the remaining differences between scans are related to local shape differences. The techniques for automatically identifying these local shape differences usually estimate high order nonlinear transforms, often with elastic matching algorithms<sup>84</sup>.

---

<sup>84</sup> G.E. Christensen, M.I. Miller, M.W. Vannier and U. Grenander. Individualizing Neuroanatomical Atlases Using a Massively Parallel Computer. *IEEE Computer*, January, 1996.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

### **7.2.4.1 Intra-patient Registration**

A common intra-patient registration method works with the concept of subsampling of the gray scale data for speed-up. Entropy calculations are performed in a histogram feature space. The algorithm is relatively fast and does not require any preprocessing of the data. The operator selects three paired landmarks and the algorithm then calculates an alignment to subvoxel accuracy. Alignment is assessed by using inherent contrast similarity to directly measure the image alignment. The algorithm requires entropy and joint entropy computation. Mutual information is defined in terms of entropy<sup>85</sup>. The first term is the entropy in the reference volume. The second term is the entropy of the part of the test volume into which the reference volume projects. It encourages transformations that project the reference volume into complex parts of the test volume. The third term, the (negative) joint entropy of the reference and test volume, contributes when they are functionally related. A histogram-based density estimate is used for the joint entropy estimation. The joint histogram computation is parallelized by dividing data into chunks, computing the histogram of each chunk, and then adding the histograms together. The joint entropy can then be calculated by a loop over the histogram. Acquisitions with different contrasts can be registered into multichannel data sets for better segmentation and visualization. Examples include image analysis in neonates (T2/PD - SPGR,) and surgical planning (MRA, SPECT, fMRI, MRI, CT)<sup>86</sup>.

### **7.2.4.2 Interpatient Registration**

In a situation where it is necessary to align two data sets of different subjects dense feature comparison turns out to be more robust than sparse feature comparison. Parallelization is used to allow the speed-up of dense feature comparisons, making the application of this technique practical in a clinical context.

---

<sup>85</sup> Wells W.M., W.E.L. Grimson, R. Kikinis, and F. A. Jolesz. 1996. Adaptive Segmentation of MRI Data. IEEE Transactions on Medical Imaging 15 (4):429-442.

<sup>86</sup> Zavaljevski A, Dhawan AP, Gaskil M, Ball W, Johnson JD, Multi-level adaptive segmentation of multi-parameter MR brain images. Comput Med Imaging Graph. 2000 Mar-Apr;24(2):87-98.

**Federation of American Scientists**  
**Biomedical Computing Requirements for HPCS**  
**Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

Segmentations of the patient scans to be aligned are generated; then a measure mismatch of alignment is generated by counting the number of voxels that don't match; then a transform that minimizes the mismatch is determined. Each scan to be registered is classified and a multiresolution pyramid of the classified scan is constructed. An initial alignment is selected as either the identity transform or the transform identified with the process described below. For each level of the pyramid, the optimum alignment is determined by minimizing the mismatch of corresponding tissue labels. Each evaluation of this mismatch can be computed in parallel.

The evaluation of a particular transform involves the comparison of aligned data with a two step process. First the moving data set is resampled into the frame of the stationary data set. Second is the voxelwise comparison of label values. Each of these steps can be parallelized by carrying out the operations simultaneously on some voxels in the frame of the stationary data set. This algorithm was initially developed for interpatient registration such as the initial alignment for template driven segmentation (TDS). TDS is used in many applications such as the quantitative analysis of MS, brain development, schizophrenia, and rheumatoid arthritis. More recently, the algorithm has been used for inpatient alignment, if a large capture range was needed.

### **7.2.5 Non-linear registration**

Non-linear registration is the set of techniques that allow the alignment of data sets that are mismatched in a nonlinear or nonuniform manner. Such misalignment can be caused by a physical deformation process, or can be due to intrinsic shape differences. For example, deformation of the brain can occur during neurosurgery with the skull is opened and CSF is drained, and when a tumor is removed. Shape differences also occur when a comparison is made between the brains of different people<sup>87</sup>. Local shape differences between data sets can be identified by finding a 3D deformation field that alters the coordinate system of one data set to maximize the similarity of local intensities with the other. Elastic matching aims to match a template, describing the anatomy expected to be present, to a particular patient scan so that the information associated with the template can be projected directly onto the patient scan on a voxel to voxel basis. The template can be an atlas of normal anatomy (deterministic or

---

<sup>87</sup> <http://spl.harvard.edu:8000/~warfield/papers/tutslides/paper>

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

probabilistic), or it can be a scan from a different modality, or it can be a scan from the same modality. The template can contain information typically found in anatomical textbooks, but unlike normal textbooks, can be linked to any form of relevant digital information.

Algorithmic improvements to speed up the processing include a multiresolution approach with fast local similarity measurement, and a simplified regularization model for the elastic membrane<sup>88</sup>. Algorithms parallelized for SMP such as low pass filter upsampling and downsampling, arithmetic operations, and solving systems of equations are typically used.

Nonlinear registration is primarily used for incremental alignment in TDS, following the linear alignment step. For an excellent discussion of nonlinear registration, see the following paper:

<http://spl.harvard.edu:8000/~warfield/papers/tutslides/paper>

## **7.2.6 Visualization**

### **7.2.6.1 Surface model generation**

To visualize the surface of structures by simulating light reflection requires generation of models by segmentation. The process consists of segmentation of the data into binary label maps and application of a surface model generation pipeline consisting of the marching cubes algorithm for triangle model generation<sup>89</sup>, followed by triangle decimation and triangle smoothing to reduce triangle count. The algorithm is parallelized by distributed computation of triangle models for each structure of a data set. Efficient triangle model generation has been used for the visual verification of segmentation procedures, visualization for surgical planning and navigation<sup>90</sup>.

### **7.2.6.2 Volume rendering**

Visualization of structures without the need for the extensive preprocessing required by the surface model approach can be done using volume rendering. This is of benefit if the structures

---

<sup>88</sup> Dengler, Joachim and Markus Schmidt. 1988. The dynamic Pyramid – A Model for Motion Analysis with Controlled Continuity. *International Journal of Pattern Recognition and Artificial Intelligence* 2 (2):275-286.

<sup>89</sup> Lorensen, W. E. and H. E. Cline, "Marching Cubes: A High Resolution 3D Surface Construction Algorithm," *Computer Graphics*, vol. 21, no. 3, pp. 163-169, July 1987.

<sup>90</sup> Najkajima, S., H. Atsumi, A.H. Bhalerao, F. Jolesz, R. Kikinis, T. Yoshimine, T. M. Moriarty, and P.E. steig. 1997. Computer-Assisted Surgical Planning for Cerebrovascular Neurosurgery. *Neurosurgery* 41 (2):403-409.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

to be visualized are constantly changing. Ray casting and shear warp algorithms are among the most popular approaches for volume rendering.<sup>91,92</sup> The algorithm is parallelized by applying the light transmission model simultaneously to different sections of the data associated with different screen pixels. Visualization of data before segmentation, visualize the magnitude of vector fields, interactive editing of volume data<sup>93</sup>.

## **8 Integrative Modeling - Tissue and Organ Modeling**

Molecules, cells, tissues--the next level of biological organization is that of the organ itself. Complex organ models attempt to take into account explicit organ geometry coupled to hydrodynamics, continuum mechanics, reaction-diffusion, radiation and discrete particle transport. Organ modeling, while still in its infancy will be the springboard for detailed modeling of the body's organs as a complete system.

Organ modeling requires coupling of models and a system integration approach to coupling the models. First the physical organization, from the molecule to the organ must be modeled. Next the integration of functional models: chemical, mechanical, electrical, metabolic, and thermal. Then the models must be extended across broader scales of time and space. For example to model the heart, the anatomy and morphology of the heart need to be represented in the geometry and structure of the continuum mechanics model. The environmental influences need to be captured in the boundary conditions. The biological processes of mass transport, growth, metabolism, energetics, motion, flow, and equilibrium must be expressed through and must

---

<sup>91</sup> Philippe Lacroute and Marc Levoy, *Fast Volume Rendering Using a Shear-Warp Factorization of the Viewing Transformation*, Proceedings of SIGGRAPH94

<sup>92</sup> Saiviroonporn, Pairash, Andre Robatino, Janos Zahajszky, Ron Kikinis, Ferenc A. Jolesz. 1998. Real Time Interactive 3D-Segmentation. *Acad Radiolo*. Vol. 5, p49-56.

<sup>93</sup> Schulze, Jurgen and Ulrich Lang. 2002. The Parallelization of the Perspective Shear-Warp Volume Rendering Algorithm. Fourth Eurographics Workshop on Parallel Graphics and Visualization.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

operate under the conservation laws for mass, energy, and momentum adopted in the model. Finally, structure-function relations need to be embodied in equations that take into account the material properties of the mechanical system.

Tissue and organ modeling generally begins with converting biology image data into computational meshes. This requires reconstruction and animation of volumetric deformable objects. In the scope of a medical application the goal is to simulate the motion and the form alteration of the organ. One current research goal is to develop better shape representations for the deformable structures found in biomedical image databases. A longer-term goal of this research is to build specialized, deformable organ shape models that learn the priors for a particular organ type. Typically digital images are used to capture the organ detail. Then, mesh grid generation software is used to analyze the data and reconstruct it into a computer model.

### **8.2.6 Organ Model Applications**

Applications include simulation of dynamic and deformable behavior of cancerous tissues and organs to be integrated in radiation dose evaluation. A computational model of the cardiovascular system is aiding researchers in understanding the fundamental biochemical, biophysical, electrical and mechanical functions of the normal heart. The model is also advancing understanding of the molecular and genetic origins of heart disease, the electrical and mechanical properties of blood flow in large and small blood vessels; and the development of potential approaches for new cardiovascular drugs. A virtual lung model, developed at the Department of Energy's Pacific Northwest National Laboratory, may help predict the impact of pollutants on respiratory systems and provide new insights into asthma, as well as other pulmonary diseases<sup>94</sup>. Using the virtual respiratory tract, PNNL scientists can analyze the influence of various factors, such as the amount of pollutants or length of exposure, on healthy versus diseased lungs by manipulating the computer model. With the model they can begin to simulate how gases, vapors and particulates may act differently within lungs of people suffering from cystic fibrosis, emphysema and asthma.

---

<sup>94</sup> <http://www.pnl.gov/news/2001/01-33.htm>

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

### **8.2.7 Heart Model Examples**

Heart models are the most advanced organ models. Two leading models are discussed here: the Peskin/McQueen model and the Cardiac Mechanics Research Group at UCSD model.

MetaCenter researchers Charles Peskin, David McQueen, and their group at the Courant Institute of Mathematical Sciences of New York University developed their heart model to help design improved artificial heart valves<sup>95</sup>. Their 3D model combines tissue and fluid mechanics models and permits the examination of the performance of modeled artificial heart valves for any of the four valves of the human heart. To solve both the fluid mechanics and elasticity problems simultaneously they use the immersed boundary method with formal second-order accuracy<sup>96</sup>. This is essentially a second-order Runge-Kutta method. The heart is modeled as a set of elastic fibers immersed in an incompressible fluid, which avoids the complexities of applying boundary conditions on the moving location of the heart walls. For greater realism, they try to make the model fibers follow the same paths as muscle and collagen fibers in the real heart muscle and valves.

High resolution is also needed to make the Reynolds number (Re) realistic. To use a realistic Re for blood flow in the heart (about 500) requires a substantial refinement of the mesh, possibly by a factor of 25 in each spatial direction. Fortunately, the flow pattern of blood in the heart is not very sensitive to the Reynolds number, and improvements in numerical methodology, such as local mesh refinement near boundaries or the use of entirely grid-free methods, may make it possible to avoid the extreme computational requirements implied by such a refinement of a uniform grid. Nevertheless, a fully satisfactory computation of blood flow in the heart will require a substantial increase in computer power. Increased computer power is needed not only to do the current computation more correctly, but also to bring in additional phenomena that are highly relevant to blood flow in the heart. Two examples are the electrical activity that

---

<sup>95</sup> McQueen, D.M., and C.S. Peskin. 2000. A three-dimensional computer model of the human heart for studying cardiac fluid dynamics. *Computer Graphics* 34:56–60.

<sup>96</sup> McQueen, D.M., and C.S. Peskin. 1997. Shared-memory parallel vector implementation of the immersed boundary method for the computation of blood flow in the beating mammalian heart. *Journal of Supercomputing* 11(3):213–236.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

coordinates and controls the heartbeat and the dynamics of the blood clotting process, which is important in evaluating the function of prosthetic cardiac valves. Models of these are being developed separately from the model of cardiac mechanics. Microscopic and macroscopic models of the clotting process are being developed by Aaron Fogelson, University of Utah. Ultimately, Peskin and McQueen hope to combine such models with their mechanical model to increase its realism and predictive power.

The model's major algorithm and approaches are:

- the immersed boundary method
- Navier-Stokes equations solved on cubic lattice through the use of FFT representations of the velocity and pressure
- fiber equations solved on Lagrangian framework
- interface equations that allow for a projection of the fiber forces onto the fluid lattice through a "smooth" delta function and interpolation of the fiber nodal velocities from fluid lattice.<sup>97</sup>

Modeling the heart presents many challenges. The heart walls and heart walls and valves move and interact with the fluid flow, both driving and responding to it. In addition, the heart muscle —is contracting and relaxing, with elastic properties that change during the contraction-relaxation cycle.

A new NPACI alpha project focused on an application for simulating blood flow in the human heart is developing a generic immersed boundary code that will run on distributed parallel machines<sup>98</sup>. The project is focused on Titanium, a parallel language and compiler, running on Blue Horizon with improved equation solvers and algorithms for handling adaptive computational grids to support —an application for simulating blood flow in the human heart. Katherine Yelick, a computer scientist at UC Berkeley who leads the project, is porting the Titanium language, which provides greater support for parallel computing, to Blue Horizon, as

---

<sup>97</sup> <http://www.npaci.edu/SAC/Collaborations/Peskin/reports/initial.profile.html>

<sup>98</sup> <http://www.npaci.edu/envision/v16.4/adaptivecomp.html>.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)*

well as porting the immersed boundary code to Titanium and developing scalable solver technology for uniform grids.<sup>99</sup> Colella is developing improved algorithms for handling adaptive computational grids, particularly for flows modeled by the immersed boundary method. Baden is developing communication support based on Kernel Lattice Parallelism (KeLP) for grid-based computation on Blue Horizon. Saltz is hardening the Titanium front-end for the Active Data Repository (ADR) storage facility to handle the immense data sets generated in the realistic simulations.

Peskin identifies intelligent adaptive mesh algorithms that will "zero in" as the flow evolves during the simulation on the computationally challenging areas where the flow is more complex, such as near the delicate valve leaflets, as a critical requirement.

As with all the above methods that use computational meshes, the Cardiac Mechanics Research Group (CMRG) [Department of Bioengineering](#) and the [Whitaker Institute for Biomedical Engineering](#) at [UCSD](#) <http://cmrg.ucsd.edu/> heart model integrates structure and function as well as theory and experiment by means of the finite element method (FEM). The group uses a prolate spheroidal coordinate system to accurately represent both the compact shape and muscle fibre architecture of the heart's muscle walls. Various parameters, derived from laboratory research, are introduced to the continuum model for comparative study. For instance, the CMRG members are working with both anatomic elements from rabbit hearts which have been histologically processed and sections revealing the orientations of the muscle fibres. A second project relates to the differences in heart muscle structure between normal and brittle-boned mice suffering from osteogenesis imperfecta (OI) because of a deficiency in the protein collagen. The finite element models showed that OI mice develop variations in the residual stresses and muscle fibre structure which constitute beneficial adaptations to the deficiency of collagen.

As far as the human heart is concerned, the CMRG investigators study the relationships between the cellular and tissue structure of the ventricular myocardium as well as the mechanical and

---

<sup>99</sup> Yelick, K., L. Semenzato, G. Pike, C. Miyamoto, B. Liblit, A. Krishnamurthy, P. Hilfinger, S. Graham, D. Gay, P. Colella, and A. Aiken. 1998. Titanium: A High-Performance Java Dialect. *Concurrency: Practice and Experience*, September-November 1998:825–36.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, khowell@fas.org*

electro-physiological function of both the intact and affected organ. In ongoing projects, the mechanisms of ventricular mechano-electric feedback, the alterations during ventricular hypertrophy, and the flow-function relations during myocardial ischemia are unveiled. In collaboration with the Cleveland Clinic Foundation and the University of Auckland in New Zealand, the researchers in the Cardiac Mechanics Research Group explore the potential of a revolutionary surgical method for patients with severe heart failure. Through the combination of computational modelling with magnetic resonance imaging, the research is to predict which patients effectively can be rescued, using surgical ventricular reduction.

In an effort that includes applications of bioinformatics, the CMRG supports *Continuity 5.5* a computational tool for continuum problems in bioengineering and physiology, especially those related to cardiac mechanics and electrocardiology research. In addition to continuum modeling, *Continuity 5.5* has facilities for least-squares fitting of parametric models to experimental measurements from diverse sources including gross anatomy, histomorphology, 3-D medical imaging, and physiological and biomechanical testing. *Continuity 5.5* is component-based using a very high-level object-oriented scripting language for component integration. Executables for *Continuity 5.5* can be downloaded free from the group's website for academic research purposes. <http://cmrg.ucsd.edu/cgi-bin/cmrg/downloads/selection.cgi>.

**Federation of American Scientists  
Biomedical Computing Requirements for HPCS  
Draft Preliminary Report**

*Contact: Kay Howell, [khowell@fas.org](mailto:khowell@fas.org)*