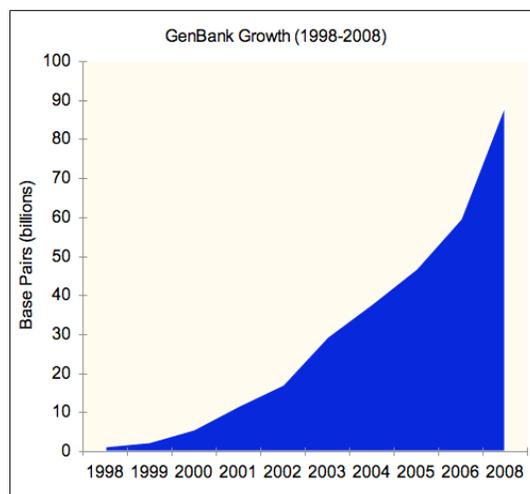


Massively Parallel Sequence Alignment: When Computational Biology Becomes I/O Biology

Wu Feng | Dept. of Computer Science | Virginia Tech | Blacksburg, VA 24060
feng@cs.vt.edu

Sequence alignment identifies regions of similarity between biological sequences. It can be used to infer functional, structural, and evolutionary relationships between sequences, assist in identifying members of gene families, and help in discovering new drugs. The standard for pairwise sequence alignment is the BLAST heuristic [1] while ClustalW [2], T-COFFEE [3], and Probcons [4] are used for multiple sequence alignment. However, despite these fast heuristic alignment algorithms, *the size of sequence databases continues to grow at a rate faster than these sequence alignment algorithms can compute on them* [5]. That is, while the speed of computational nodes double roughly every 24 months, database sizes double every 12 months. Consequently, parallel tools such as mpiBLAST [6-9] were introduced, but even these will not be enough to handle the oncoming onslaught of genomic data due to the advent of metagenomics [10] and next-generation sequencers from Illumina and Applied Biosystems [11-13], which will further exacerbate the problem by producing as many sequences in one week as it took the entire GenBank database to accumulate over its 27-year lifetime [14]. Finally, *dealing with this amount of data, while still computationally intensive, will shift the performance bottleneck from computation to I/O, i.e., input/output*. Examples of the above issues can be demonstrated across a myriad of applications including *finding missing genes in genomes, real-time pathogen detection, and personalized genomics*, including genome-wide, single-nucleotide polymorphic (SNP) analysis of the human genome for disease associations. Thus, investments in massively parallel sequence alignment and its associated areas will be critical to the plethora of biological applications that are downstream from sequence alignment.



Below we briefly address the challenges and importance of tackling *one* of the above problems: *finding missing genes in genomes*, a new area of research that has the potential of enabling a potpourri of biological research, thus making seed investments invaluable, given the expected payoff, e.g., efficiently identifying enzymes to enable an industrial process. Finding missing genes, in collaboration with Prof. Joao Setubal who proposed the problem, can provide a more complete picture of the capabilities of the organism in question. If the organism is a pathogen, then one may be better able to control the disease it causes; if the organism is beneficial, one may be able to understand better its metabolism and hence improve its “efficiency.” In all cases, it will improve our knowledge of the repertoire of protein-coding genes found in nature, and that in itself, can lead to other discoveries, e.g., an enzyme that can be used in some industrial process. In the case of the ORNL Bioenergy Science Center, for instance, it means more readily identifying the necessary enzymes to efficiently convert biomass sugars into hydrogen or electrical energy [15, 16]. Thus far, *we have already uncovered hundreds of missing genes* [17], a process that is outlined below.

In 2007, conventional wisdom espoused that *finding missing genes* in 567 microbial genomes was computationally infeasible because it entailed $O(10^{15})$ massively parallel sequence alignments, followed by significant post-processing. However, by leveraging my open-source mpiBLAST cybertool (<http://www.mpiblast.org/>) [6-9], I led a team of 15 interdisciplinary researchers from 7 institutions around the world and developed additional software cybertools to integrate a set of distributed supercomputers, totaling 12,000+ processor cores and approximately 0.2 petaflops in performance, to

tackle the massively parallel sequence alignment portion of finding missing genes in genomes and to store approximately a petabyte of output in Tokyo, Japan, the only place where I could identify sufficient storage [18]. (The post-processing details of finding missing genes can be found in [17].)

The raw prototyping of our cybertools took on the order of two months for the team to develop while the massively parallel sequence alignment would have taken approximately *three years* to complete. Why so long? While leveraging large-scale parallelism dramatically dropped the computation time, each parallel compute node generated enough output data to cause the performance bottleneck to shift dramatically from computation to I/O, i.e., computational biology to I/O biology. To address this issue, we created *ParaMEDIC: Parallel Metadata Environment for Distributed I/O and Computing* to reduce the task from over 156 weeks to a mere 2 weeks [18-20].

However, the above process took a heroic effort of personnel resources and cyberinfrastructure resources to achieve. This is a problem that is not specific to just finding missing genes, it also applies to a host of other applications that are downstream from sequence alignment, such as those noted earlier, i.e., genome-wide, single-nucleotide polymorphic (SNP) analysis, real-time pathogen detection, personalized genomics, and so on. Furthermore, I have only talked about pairwise sequence alignment here; the problems in multiple sequence alignment are even more daunting.

Accelerating the Computation of Long-Range Interactions: Towards Rational Drug Design

In molecular modeling, whether classical or quantum, long-range interactions are difficult to compute *accurately* because the interactions between all pairs of points must be computed. (This general problem also has direct applicability to cosmology, e.g., <http://www.astrogrape.org/>, which has won numerous Gordon Bell Awards at ACM/IEEE Supercomputing.)

We are specifically looking to speed-up implicit solvent molecular dynamics with a newly developed HCP algorithm, which exploits the natural partitioning of biomolecules into its constituent components to speed-up the computation of pairwise electrostatic interactions with a limited and controllable impact on accuracy [21]. For large systems, the $O(N \log N)$ complexity of HCP delivers up to 3 orders of magnitude in speed-up [21] relative to the reference exact $O(N^2)$ computation. If HCP could be combined with our current 3 orders-of-magnitude speed-up of electrostatic interactions on the GPU [22], the resulting *million-fold* speed-up could revolutionize the field of molecular modeling.

These long-range interaction calculations can, in turn, contribute in the process of rational drug design. Because one must find a small molecule that blocks the function of a particular enzyme, e.g., the viral protein responsible for AIDS, the above calculations, when appropriately mapped and then coupled with understanding the precise 3D structure of that protein, can lead to successful rational drug design, as done with the drug Sustiva, one of the drugs that stopped the AIDS epidemic in the U.S., i.e., part of the anti-retro viral cocktail.

In closing, while investing in the above areas will enable researchers to conduct their research faster, it more importantly empowers researchers across disciplines to tackle problems previously viewed as infeasible or that require heroic efforts and significant domain-specific expertise to solve. In short, it will enable researchers with the *power to attempt problems previously viewed as infeasible* as well as the *power to run "What if?" scenarios at will*. In the long term, we hope to commoditize such endeavors, but for now, simply identifying and securing the needed petascale resources, both in terms of personnel and cyberinfrastructure, to tackle the above problems remains a significant challenge.

References

- [1] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic Local Alignment Search Tool," *J. Molecular Biology*, 215(3):403-410, 1990.
- [2] J. Thompson, D. Higgins, and T. Gibson, "ClustalW: Improving the Sensitivity of Progress Multiple Sequence Alignment ...," *Nucleic Acids Research*, 22(22):4673-80, Nov. 1994.
- [3] C. Notredame, D. Higgins, and J. Heringa, "T-Coffee: A Novel Method for Multiple Sequence Alignments," *J. Molecular Biology*, 302:205-217, 2000.
- [4] C. Do, M. Mahabhashyam, M. Brudno, and S. Batzoglou, "PROBCONS: Probabilistic Consistency-based Multiple Sequence Alignment," *Genome Research*, 15: 330-340, 2005.
- [5] F. Meyer, "Genome Sequencing vs. Moore's Law: Cyber Challenges for the Next Decade," *CTWatch Quarterly*, 2, 2006.
- [6] J. Archuleta, E. Tilevich, and W. Feng, "A Maintainable Software Architecture for Fast and Modular Bioinformatics Sequence Search," *IEEE International Conference on Software Maintenance*, 144-153, October 2007.
- [7] A. Darling, L. Carey, and W. Feng. The Design, Implementation, and Evaluation of mpiBLAST. *4th International Conference on Linux Clusters*, **Best Paper**, June 2003.
- [8] W. Feng and A. Darling, "mpiBLAST: A High-Speed Software Catalyst for Genetic Research," ***R&D 100 Award***, 2004.
- [9] H. Lin, P. Balaji, R. Poole, C. Sosa, X. Ma, and W. Feng, "Massively Parallel Genomic Sequence Search on the Blue Gene/P Architecture," *ACM/IEEE SC '08 (Supercomputing)*, Nov. 2008.
- [10] National Research Council, *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. National Academy of Sciences, 2007.
- [11] 454 Life Science, Products and Solutions. <http://www.454.com/products-solutions/system-features.asp>.
- [12] Illumina Inc, "Illumina Presents Development Roadmap for Scaling its Genome Analyzer," <http://www.reuters.com/article/pressRelease/idUS132680+05-Feb-2009+BW20090205>.
- [13] J. Perkel, "Sanger Who? Sequencing the Next Generation," *Science*, 2009.
- [14] D. Benson, I. Karsch-Mizrachi, D. Lipman, J. Ostell, and D. Wheeler, "GenBank," *Nucleic Acids Research*, 36 (Database issue), January 2008.
- [15] X. Ye, Y. Wang, R. Hopkins, M. Adams, B. Evans, J. Mielenz, and P. Zhang, "Spontaneous High-Yield Production of Hydrogen from Cellulosic Materials and Water Catalyzed by Enzyme Cocktails," *ChemSusChem*, 2(2):149-152, 2009.
- [16] P. Zhang, B. Evans, J. Mielenz, R. Hopkins, and M. Adams, "High-Yield Hydrogen Production from Starch and Water by a Synthetic Enzymatic Pathway," *PLoS ONE*, 2(5):e456, 2007.
- [17] A. Warren, J. Archuleta, W. Feng, and J. Setubal, "Missing Genes in the Annotation of Prokaryotic Genomes," In preparation.
- [18] P. Balaji, W. Feng, J. Archuleta and H. Lin, "ParaMEDIC: Parallel Metadata Environment for Distributed I/O and Computing," *ACM/IEEE SC '07 (Supercomputing)*, **Storage Challenge Award**, November 2007.
- [19] P. Balaji, W. Feng, and H. Lin, "Semantic-Based Distributed I/O with the ParaMEDIC Framework," *17th ACM/IEEE International Symposium on High-Performance Distributed Computing*, 175-184, June 2008.
- [20] P. Balaji, W. Feng, H. Lin, J. Archuleta, S. Matsuoka, A. Warren, J. Setubal, E. Lusk, R. Thakur, I. Foster, D. Katz, S. Jha, K. Shinpaugh, S. Coghlan, and D. Reed. Distributed Data I/O with ParaMEDIC: Experiences with a Worldwide Supercomputer. *International Supercomputing Conference (ISC)*, **Best Paper Award**, June 2008.
- [21] R. Anandakrishnan and A. Onufriev, "An N log N Approximation Based on the Natural Organization of Biomolecules for Speeding up the Computation of Long Range Interactions," *J. Computational Chemistry*, 2009.
- [22] R. Anandakrishnan, T. Scogland, A. Fenley, J. Gordon, W. Feng, and A. Onufriev, "Accelerating Electrostatic Surface Potential Calculation with Multiscale Approximation on Graphics Processing Units," Technical Report, Virginia Tech, 2009.

State of the art, Opportunities in Biology at the Extreme Scale of Computing, mini-white paper, Ed DeLong

For the purposes of this document, ‘metagenomics’ is defined as cultivation-independent genomic analysis of microbial assemblages or populations. While still in its infancy, metagenomics has already contributed significantly to our knowledge of the genomic structure, population diversity, gene content, and composition of naturally occurring microbial assemblages. In low complexity populations metagenomic studies have led to the assembly of near complete genomes from dominant genotypes(13), and have provided composite genomic representations of dominant populations(1, 7). Despite the large datasets now available however, high allelic variation in microbial populations, high species richness, and relatively even representation among species, still render whole genome assemblies of individual genotypes impractical, given current sequencing and assembly technologies(10, 14, 15). Given appropriate data scale-up enabled by 2nd and 3rd gen sequencing technologies, appropriate computation algorithms and petaflops of computing power will certainly be required to address this problem on large scales.

A major challenge in emerging metagenomic, ‘metatranscriptomic’ and ‘metaproteomic’ studies are the sheer size of the datasets, and the new methods and tools and computational infrastructure that is needed to deal with their magnitude. Size matters. These exponentially growing datasets raise new challenges with respect to data management, computational resources, sampling and analytical strategies, and database architectures – no currently met by existing algorithms or computational infrastructures and capacities. The need to establish standards for metadata submission and reporting, so that primary sequence data can be related across relevant environmental parameters is clear. The Genomic Standards Consortium (GSC) are promoting schemes reminiscent of the MIAME standards for microarray data (<http://www.mged.org/Workgroups/MIAME/miame.html>), that would capture metadata associated with genomes (Minimum Information about a Genome Sequence, MIGS), and

metagenomic data (Minimum Information about a Metagenome Sequence, MIMS)(3, 6). For archived datasets, such metadata field standardization and reporting will be critical. As mentioned above, we are entering a new era in microbial ecology and biology, that will increasingly employ high-throughput sequencing data as an analyte in experimental protocols. Coordination of experimental reports from such inquiries will be important, and MIAME-like standards for such reporting (Minimum Information about a high-throughput SeQuencing Experiment – MINSEQE) have recently been proposed as well (<http://www.mged.org/minseqe/>). Even “simple” annotation, archiving, and accessing of the new sequence data types and experiments, along with associated and relevant metadata, poses significant challenges for the biological community. These challenges are now beginning to be addressed by the development of new types of metagenomic databases(8, 9, 11), analytical strategies and statistical approaches. But still, the tools and compute power for actually analyzing and comparing the data is vastly outstripped by the data. Iterative “all against all” blast runs, for example, are out of the question with current compute resources.

There are many new and evolving methodologies that extend metagenomic approaches further along the hierarchy of biological organization - notably transcriptomics and proteomics are being applied successfully to the study of complex natural microbial populations. Development of microbial community transcriptomic methods is enabling a new research agenda in microbial ecology, that utilize sequence data as an analyte in experimental field studies(5, 12). The approach enables the measurement of microbial assemblage gene expression in microcosms, mesocosms or natural samples, as a function of environmental variability over time. The environmental variation examined can be natural (for example, tracking changes in gene expression as a function of the diel cycle), or applied (for example, monitoring changes in gene expression following nutrient emendation). By tracking genes responsive to specific environmental perturbations, it should soon be possible to track environmental perturbations that are first observable as changes in gene expression in resident microbial populations, but that later may lead to shifts in community composition. Quantifying the variability and kinetics of gene expression in natural assemblages has potential to provide

a fundamentally new perspective on microbial community dynamics. Can expression patterns provide clues as to the functional properties of hypothetical genes? What are the key community responses to natural or anthropogenic environmental perturbation? Are there fundamental community-wide regulatory responses common to disparate taxa? Are certain taxa or metabolic paths more or less responsive to particular environmental changes? Are specific changes in gene expression indicative of downstream changes in community composition?

In similar ways, ‘metaproteomics’ adds dramatically to the complexity of datatypes, observational and experimental scenarios, and computational challenges that need to be met. One can easily imagine complex datasets with rich environmental metadata for which phylogenetic survey, metagenomic, transcriptomic, and proteomic data also exists. Extending such data across different spatial and temporal gradients and scales, is well within the reach of current technologies. Computation and data management infrastructures however, for organizing and analyzing such data do not currently exist. Indeed, complex physical models like the MIT Global Circulation model, coupled with biological ecological and evolutionary modeling, are already beginning to be used(4) – requiring large amounts of compute resources to run just the physical models, not to mention the evolutionary modeling.

Efficient bioinformatics management and analytical practices will not be a panacea for the larger challenge of describing microbial biology at an ecosystem level. There still exists a significant mismatch with respect to integrating “bottom up” reductionist molecular approaches, with “top down” integrative ecosystems analyses. Molecular datasets are often gathered in massively parallel ways, but acquiring equivalently dense microbial and biogeochemical process data(2) is not currently as feasible. This ‘impedance mismatch’ (e.g., inadequate (or excessive) ability of one system to accommodate the input from another), is one of the larger hurdles that will have to be overcome for more realistic, integrative analyses that interrelate datasets spanning from genomes to biomes.

The need for more advanced computational infrastructures for such analyses is evident and urgent.

1. **Allen, E. E., G. W. Tyson, R. J. Whitaker, J. C. Detter, P. M. Richardson, and J. F. Banfield.** 2007. Genome dynamics in a natural archaeal population. *Proc Natl Acad Sci U S A* **104**:1883-8.
2. **Anderson, R., D. Archer, U. Bathmann, P. Boyd, K. Buesseler, P. Burkill, A. Bychkov, C. Carlson, C. T. Chen, S. Doney, H. Ducklow, S. Emerson, R. Feely, G. Feldman, V. Garcon, D. Hansell, R. Hanson, P. Harrison, S. Honjo, C. Jeandel, D. Karl, R. Le Borgne, K. Liu, K. Lochte, F. Louanchi, R. Lowry, A. Michaels, P. Monfray, J. Murray, A. Oschlies, T. Platt, J. Priddle, R. Quinones, D. Ruiz-Pino, T. Saino, E. Sakshaug, G. Shimmield, S. Smith, W. Smith, T. Takahashi, P. Treguer, D. Wallace, R. Wanninkhof, A. Watson, J. Willebrand, and C. S. Wong.** 2001. A new vision of ocean biogeochemistry after a decade of the Joint Global Ocean Flux Study (JGOFS). *Ambio*:4-30.
3. **Field, D., G. Garrity, T. Gray, N. Morrison, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, M. J. Allen, S. V. Angiuoli, M. Ashburner, N. Axelrod, S. Baldauf, S. Ballard, J. Boore, G. Cochrane, J. Cole, P. Dawyndt, P. De Vos, C. DePamphilis, R. Edwards, N. Faruque, R. Feldman, J. Gilbert, P. Gilna, F. O. Glockner, P. Goldstein, R. Guralnick, D. Haft, D. Hancock, H. Hermjakob, C. Hertz-Fowler, P. Hugenholtz, I. Joint, L. Kagan, M. Kane, J. Kennedy, G. Kowalchuk, R. Kottmann, E. Kolker, S. Kravitz, N. Kyrpides, J. Leebens-Mack, S. E. Lewis, K. Li, A. L. Lister, P. Lord, N. Maltsev, V. Markowitz, J. Martiny, B. Methe, I. Mizrachi, R. Moxon, K. Nelson, J. Parkhill, L. Proctor, O. White, S. A. Sansone, A. Spiers, R. Stevens, P. Swift, C. Taylor, Y. Tateno, A. Tett, S. Turner, D. Ussery, B. Vaughan, N. Ward, T. Whetzl, I. San Gil, G. Wilson, and A. Wipat.** 2008. The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* **26**:541-7.
4. **Follows, M. J., S. Dutkiewicz, S. Grant, and S. W. Chisholm.** 2007. Emergent biogeography of microbial communities in a model ocean. *Science* **315**:1843-6.
5. **Frias-Lopez, J., Y. Shi, G. W. Tyson, M. L. Coleman, S. C. Schuster, S. W. Chisholm, and E. F. DeLong.** 2008. Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci U S A* **105**:3805-10.
6. **Garrity, G. M., D. Field, N. Kyrpides, L. Hirschman, S. A. Sansone, S. Angiuoli, J. R. Cole, F. O. Glockner, E. Kolker, G. Kowalchuk, M. A. Moran, D. Ussery, and O. White.** 2008. Toward a standards-compliant genomic and metagenomic publication record. *Omics* **12**:157-60.
7. **Hallam, S. J., K. T. Konstantinidis, N. Putnam, C. Schleper, Y. Watanabe, J. Sugahara, C. Preston, J. de la Torre, P. M. Richardson, and E. F. DeLong.** 2006. Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proc Natl Acad Sci U S A* **103**:18296-301.

8. **Markowitz, V. M., N. N. Ivanova, E. Szeto, K. Palaniappan, K. Chu, D. Dalevi, I. M. Chen, Y. Grechkin, I. Dubchak, I. Anderson, A. Lykidis, K. Mavromatis, P. Hugenholtz, and N. C. Kyrpides.** 2008. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* **36**:D534-8.
9. **Meyer, F., D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards.** 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**:386.
10. **Rusch, D. B., A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y. H. Rogers, L. I. Falcon, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Neilson, R. Friedman, M. Frazier, and J. C. Venter.** 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**:e77.
11. **Seshadri, R., S. A. Kravitz, L. Smarr, P. Gilna, and M. Frazier.** 2007. CAMERA: a community resource for metagenomics. *PLoS Biol* **5**:e75.
12. **Shi, Y., G. W. Tyson, and E. F. DeLong.** 2009. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459**:266-9.
13. **Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield.** 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**:37-43.
14. **Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Neilson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith.** 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**:66-74.
15. **Yooseph, S., G. Sutton, D. B. Rusch, A. L. Halpern, S. J. Williamson, K. Remington, J. A. Eisen, K. B. Heidelberg, G. Manning, W. Li, L. Jaroszewski, P. Cieplak, C. S. Miller, H. Li, S. T. Mashiyama, M. P. Joachimiak, C. van Belle, J. M. Chandonia, D. A. Soergel, Y. Zhai, K. Natarajan, S. Lee, B. J. Raphael, V. Bafna, R. Friedman, S. E. Brenner, A. Godzik, D. Eisenberg, J. E. Dixon, S. S. Taylor, R. L. Strausberg, M. Frazier, and J. C. Venter.** 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* **5**:e16.

*** What specific problem could be attacked and solved with the application of sustained multiple petaflops of computing power? What progress could be obtained on the problem at roughly the 10, 100, and 1000 petaflops levels of sustained performance?**

In random order:

- Understanding the brain – artificial neural networks, true autonomy (e.g., [6-9, 2])
- Genomics (genotype – phenotype)
- Proteomics (including protein folding, de novo protein design, e.g., [3-5, 1])
- Drug design
- Simulation of cellular pathways
- Protein expression
- Chemical reaction networks

*** Is the problem one of the “top 10” problems for the scientific discipline, independent of computing? Who would constitute the community of scientists and/or engineers that would enthusiastically address the problem? What would be the degree of international potential participation?**

Brain function, true autonomy, protein folding/design and cellular pathways are certainly among the top 10 problems, even independent of computing. The community of scientists that would enthusiastically address the problem comprises computational biologists, software architects, computer scientists (in particular also working in high-dimensional optimization), mathematicians, physicists (for modeling), and biologists. There is a high degree of international potential participation as these are overarching problems that are of broad international interest.

*** How is the use of petascale computational modeling and simulation irreplaceable in answering this question? Does it augment existing techniques or replace them? Is there history of large-scale computation being the preferred approach for this problem?**

Especially in the protein folding and design field there is a history of large-scale computing being the preferred approach, as more relevant (larger and more complex) molecules can be tackled with more computational resources (e.g., [3-5]). However, one has to exercise caution in developing codes that are actually scalable and thus can take advantage of petascale computing (e.g., Simulated Annealing in the optimization community allows for a linear speed-up [1]).

*** Why are the other techniques (e.g., experiments/observation, more traditional theory) that could answer these questions not satisfactory? Is it even feasible to consider other techniques?**

Because of the inherent complexity these problems cannot be tackled or solved in a timely manner with just experimental and traditional theory approaches. Rather computational modeling, multi-parameter optimization, and data mining can integratively inform and guide the actual experimentation in the laboratory. That way the experiment validates the computational models and predictions.

*** What is the current status of the computing tools for the work being proposed:**

mathematical models, algorithms, software, and data analysis tools? What is the largest scale to date that codes have been run? (e.g. 1,000, 10,000, 100,000 cores) Are there existing code teams working on codes for this problem area, or is this a new area that would need seed investments?

As for multi-dimensional optimization (see also “Stochastic Optimization Framework” below), these have been run on $O(1000)$ (i.e., thousands) of CPUs. The optimization community works on parallelizing these codes, or making them at least distributable (see also [1]).

*** What experimental and observational data is there available to validate the codes? Is the validation method well established?**

In the case of protein design, the computationally proposed proteins can subsequently be synthesized in the lab to test out the folding properties and biological activities (e.g., [3-5]).

*** What are the missing pieces in the areas of mathematical models, algorithms, software required to solve the problem? How would you rank them in terms of importance, cost, and risk?**

- Mathematical models, numerical methods, and simulations that are scalable.

Some of the computational tools and simulation/numerical techniques do exist and are well-established. Others, such as Stochastic Optimization Frameworks [1] using Simulated Annealing [10, 11, 1], Genetic/Evolutionary Algorithms [12, 13], and Genetic Programming [14] for high-dimensional optimization (Fink et al., 2008, 2009), need to be much more introduced into the respective special interest communities and subsequently interfaced to the respective problems at hand. Thus one of the bigger challenges lies in the dialog between interdisciplinary scientists to familiarize each other with the tools and problems, and to jointly generate/create appropriate/high-fidelity models and solutions that address the key problems in biology.

Modeling Scale

Another question of importance is when it is appropriate (matter of accuracy) and sufficient (matter of computational resources/time) to use micro-, meso-, or macroscopic modeling. Akin to the quantum mechanical description at small scales and a mechanical description at larger scales, it is important to judge the level of granularity that has to be applied to understanding the respective aspects of biological problems. This has direct implications on what computation resources or methods need to be employed: workstation, cluster computers, super computers, distributed computing, grid computing, lattice computations, simulation tools such as Mathematica and Matlab. In general, full-blown, quantitative spatio-temporal simulations should be aimed at. The techniques for this are well-established (e.g., lattice simulations with diffusion dynamics). Furthermore, for microscopic simulations game theory should be applied as well to study and simulate individual agent interactions, which will yield both quantitative and qualitative as well as spatio-temporal results. For large-scale simulations, simulation techniques from astrophysics (star-simulations, adaptive grid/mesh calculations) and fluid dynamics should be looked at and adopted. The actual challenge resides in the capability to model the respective

systems appropriately. This challenge can be overcome by putting together interdisciplinary teams of scientists, e.g., biologists, physicists, mathematicians.

Complexity related questions to be addressed:

There is a need for a comprehensive definition of “complexity”, especially as it pertains to microbes and interfaces. Related to the complexity question is the question of “reducibility” of complex systems, i.e., can the complex behavior of an observed biological system (e.g., microbial interface) be broken down into its (agent-based) constituents and be fully understood at the agent/agent interaction level, or are there synergistic effects that make it irreducible. This has direct implications as to how to treat the system, i.e., microscopically or macroscopically. Related to this question of “reducibility” is the question of “downward causation” in complex systems: can global indicators/objectives (e.g., climate change, hazardous waste, biofuel production rate) influence the agent level from which the complex system emerged?

References:

1. Fink W, Stochastic Optimization Framework (SOF) for Computer-Optimized Design, Engineering, and Performance of Multi-Dimensional Systems and Processes; *Proc. SPIE*, Vol. 6960, 69600N (2008); DOI:10.1117/12.784440
2. Fink W (2009) Autonomous Self-Configuration of Artificial Neural Networks for Data Classification or System Control; *Proc. SPIE*, Vol. 7331, 733105 (2009); DOI:10.1117/12.821836
3. B.I. Dahiyat, S.L. Mayo, De novo protein design: fully automated sequence selection, *Science* 278(5335):82-7, 1997.
4. J. Desmet, M. de Maeyer, B. Hazes, I. Lasters, The dead-end elimination theorem and its use in protein side-chain positioning, *Nature*, 356, 539-542, 1992.
5. C.A. Voigt, D.B. Gordon, S.L. Mayo, Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design, *J Mol Biol* 299(3):789-803, 2000.
6. McCulloch W, Pitts W, A logical calculus of the ideas immanent in nervous activity, *Bulletin of Mathematical Biophysics*, 7:115-133, 1943.
7. Hopfield JJ, Neural networks and physical systems with emergent collective computational abilities *Proc. Natl. Acad. Sci. USA* 79 2554–8, 1982.
8. Hertz J, Krogh A, Palmer RG, *Introduction To The Theory Of Neural Computation*, Lecture Notes Volume I, Addison-Wesley Publishing Company, 1991.
9. Müller B, Reinhardt J, *Neural Networks: An Introduction*, Springer, Berlin Heidelberg New York, 1990.
10. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E, Equation of State Calculation by Fast Computing Machines, *J. of Chem. Phys.*, 21, 1087 – 1091, 1953.
11. Kirkpatrick S, Gelat CD, Vecchi MP, Optimization by Simulated Annealing, *Science*, 220, 671 – 680, 1983.
12. Holland JH, *Adaptation in Natural and Artificial Systems*, The University of Michigan Press, Ann Arbor, Michigan, 1975.
13. Goldberg DE, *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, 1989.
14. Koza JR, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, Cambridge, MA: The MIT Press, 1992.

Establishing Ground Truth

George M. Garrity, Michigan State University

As a systematic biologist, I view myself as an interested party who works on the periphery of computational biology, benefitting from the advances in the field, but not necessarily contributing directly to core activities. Therefore, I will defer to others in addressing the questions that were posed by the conveners of this workshop that focus on problems that can only be solved by high-end computing. Instead, I would like to offer some thoughts on opportunities and challenges that are clearly visible from my vantage point, that will benefit from developments in high-end computing, and will help to build a new foundation for the broad field of microbiology.

It is a given that the volume and complexity of sequence data will continue to grow super-linearly for the foreseeable future, as new computational methods are applied to answer the “big questions” in biology. It is also a given that the quality of associated data and metadata will continue to be a major source of variability in our analyses because much of that data is gathered from secondary sources rather than produced *de novo*. Such data are more likely to contain errors of commission and omission and to be affected by semantic ambiguity and hidden biases. Yet, there is a tendency for these data to flow into our analyses, to affect the interpretation of our models in unpredictable ways, and to ultimately flow back into the literature. Perhaps it is because producing non-sequence data is more labor intensive or does not scale well or that interpretation requires human intervention that justifies such actions. But, what is the real cost?

The impact of semantic ambiguity in biological data has been previously noted with respect to identifiers [1] and biological names [2]. Such ambiguity confounds the accurate and complete retrieval of biological data from public and private databases and from the biological literature. This is especially true for biological names as taxonomic information is often misinterpreted or the source organisms misidentified. Bortolus [3] provides some interesting examples of error cascades in the biological sciences caused by this problem. While his remarks were aimed at field ecologists, they apply equally well to computational biologists, modelers, and system biologists. As he points out, such errors impact our knowledge of nature and have significant socioeconomic costs in some cases. Such errors are also likely to scale well, even if the methods used to produce the data do not. Laurin [4] and Hillis [5] provide a glimpse of yet another challenge that looms on the near horizon. Application of the Phylocode system of nomenclature to plants and animals will add yet another layer of complexity to mining public databases and the literature. The resulting methodological and theoretical bias it will introduce will need to be factored into the interpretation of biological data in the future.

This is not unfamiliar territory to those who have carried out “large-scale” phylogenetic, taxonomic, or ecological analyses in the past. Incorrectly labeled data and data derived from incorrectly identified samples remain common and will continue to confound naive users of public databases and the literature. Tools and techniques to detect and visualize such discrepancies have been built by a number of workers and could be useful as components of analytical pipelines, data submission routines or as value added services. So

too, would be a thorough cleaning of the public data sets if there were some way to reliably automate the task and to establish persistent links to metadata so that it would be available on demand. Implementation of strong policies on data quality that augment data sharing policies could ensure that this problem would not be exacerbated in the future. Absent this, the benefits of high-end computing will not be fully realized by the larger community.

One alternative to such an approach is to create and maintain authoritative reference sets of gene and genome sequences, derived from taxonomic type strains and other reference strains of importance, and to persistently link those reference sequences to validated phenotypic, physical, and geographic metadata and that is delivered in a highly structured, standards compliant form [6,7]. *The Genomic Encyclopedia of Bacteria and Archaea* (<http://www.jgi.doe.gov/programs/GEBA/>) is an outstanding example of an international collaboration to produce such high value benchmarks and it provide much of the ground truth for the next generation of large-scale phylogenetic models of the bacteria and archaea.

1. Clark T. Identity and interoperability in bioinformatics. *Brief Bioinform* 2003;4(1):4-6 PMID:12715829
2. Garrity GM, Lyons C. Future-proofing biological nomenclature. *OMICS* 2003;7:31-31
3. Bortolus A. Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. *Ambio* 2008;37(2):114-118 PMID:18488554
4. Laurin M. The splendid isolation of biological nomenclature. *Zoologica Scripta* 2008;37(2):223-233
5. Hillis DM. Constraints in naming parts of the Tree of Life. *Mol Phylogenet Evol* 2007;42(2):331-338 PMID:16997582 doi:S1055-7903(06)00308-3 [pii]
10.1016/j.ympev.2006.08.001
6. Garrity GM, Field D, Kyrpides N, Hirschman L, Sansone SA, Angiuoli S, Cole JR, Glockner FO, Kolker E, Kowalchuk G, *et al.* Toward a standards-compliant genomic and metagenomic publication record. *OMICS* 2008;12(2):157-160 PMID:18564916 doi:10.1089/omi.2008.A2B2
7. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, *et al.* The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* 2008;26(5):541-547 PMID:18464787 doi:nbt1360 [pii]
10.1038/nbt1360

DOE Extreme Computing Workshop: Opportunities in Biology at the Extreme Scale of Computing

Breakout Session: Populations, Communities, Ecosystems and Evolutionary Dynamics: Genomics and Metagenomics

Elebeoba May, PhD
Principal Member Technical Staff
Sandia National Laboratories

What specific problem could be attacked and solved with the application of sustained multiple petaflops of computing power?

Accurate modeling and simulation of multiscale biological phenomena, such as quorum sensing-mediated biofilm formation and microbial communities interaction with the plant rhizosphere, requires advanced computing power and mathematically robust methods that can incorporate the rapidly varying (seconds) dynamics of intracellular pathways and the slower (hours to days) time steps involved in the emergence of observable multicellular phenotypes. Increased computational and algorithmic resources will enable accurate and predictive simulation that captures the interplay between spatial interactions and temporal signal propagation. As we move from microbial colonies to multicellular organisms, the number of intracellular, cellular, and extracellular components grows exponentially. With an estimated trillions of cells in the human body, not counting the estimated three pounds of bacteria in our digestive tract (<http://www.usnews.com/articles/science/plants-animals/2008/04/08/microbes-to-people-without-us-youre-nothing.html>), the arrival of exaflop computing can bring the molecular simulation of multicellular organisms closer to reality. Exascale computing can also enable simulation of microbial communities present in the soil (~1 billion bacteria per gram of soil, ~1 billion microbes per liter of sea water). The realization of predictive multiscale simulation of organisms and microbial communities must capture not only the cellular dynamics among billions and trillions of cells, but also the thousands of components in each cell that ultimately determine individual cellular response which results in an observable phenotypic response. This is a present challenge in computational biology that will directly benefit from exascale computing.

What is the current status of the computing tools for the work being proposed: mathematical models, algorithms, software, and data analysis tools?

Several approaches for multiscale simulation of biological systems are emerging. In addition to methods such as agent-based modeling, multiscale simulation approaches include:

- Coupled ordinary differential equations (ODE) and partial differential equation (PDE) methods (Dockery and Keener, 2001; Chopp et al., 2003)
- Kinetic Monte Carlo (KMC) methods (Shrout et al., 2006)
- Cellular Potts model (Jiang et al., 2005)

To take advantage of the multiscale potential of exascale systems, further development of computational methods that enable coupling of intracellular, extracellular, and multicellular level reaction-diffusion model is needed. Coupling ODE-based intracellular models implemented in BioXyce large-scale, parallel biochemical circuit simulator to SPAARKS, a KMCcode can enable the simulation of multiscale spatiotemporal dynamics of microbial systems (May and Schiek, 2009; Slepoy, et al. 2008).

Empirical data for multiscale phenomena is increasing as technology enables observations at lower length and time scales. Genomic, proteomic, and metabolomic data as well as physiological observations enable the reconstruction of stimulus-dependent microbial behavior. Advances in micro/nanosystems are enabling the observation of intracellular events at the single-cell level. This multi-faceted dataset will facilitate calibration of mathematical models and simulations of biological systems. However they cannot replicate all possible perturbations a given system may encounter, this is the realm and value of simulation-enabled science.

What are the missing pieces in the areas of mathematical models, algorithms, software required to solve the problem?

Further advances in computational methods to enable exascale computational biology are needed. These include:

- Development of computational methods and code partitioning approaches to couple various simulation methods into a single multiscale platform.
- Challenges in coupling stochastic and deterministic simulators
- Methods for automatic run-time coarse/fine-graining various parts of model
- Methods that enable the growing and shrinking of the computational model as the biological system grows/shrinks through cell division/death.
- Given the inherent complexity of the phenomenon, the multiscale nature of the data, and the large amount of data used and produced in the simulation, methods for large-scale and advanced data visualization capabilities are needed.

References

D. L. Chopp, M. J. Kirisits, B. Moran, and M. R. Parsek. The dependence of quorum sensing on the depth of a growing biofilm. *Bulletin of Mathematical Biology*, 65:1053–1079, 2003.

J. D. Dockery and J. P. Keener. A mathematical model for quorum sensing in *Pseudomonas aeruginosa*. *Bulletin of Mathematical Biology*, 63:95–116, 2001.

E. May, R. Schiek, BioXyce: An engineering platform for the study of cellular systems. *IET Systems Biology Journal*, 3(2):77-89, March 2009.

J. D. Shrout, D. L. Chopp, C. L. Just, M. Hentzer, M. Givskov, and M. R. Parsek. The impact of quorum sensing and swarming motility on *Pseudomonas aeruginosa* biofilm formation is nutritionally conditional. *Molecular Microbiology*, 62(5):1264–1277, 2006.

A. Slepoy, A. P. Thompson, and S. J. Plimpton. A Constant-Time Kinetic Monte Carlo Algorithm for Simulation of Large Biochemical Reaction Networks. *J Chem Phys*, 2008.

Y. Jiang, J. Pjesivac-Grbovic, C. Cantrell, and J. P. Freyer. A Multiscale Model for Avascular Tumor Growth. *Biophysical Journal*, 89:3884–3894, December 2005

Metagenome analysis for the next decade

The Metagenomics RAST server {Meyer, 2008 #337} is a public, web-based resource for the analysis and comparison of large shotgun metagenomics data sets. Together with a next-generation sequencing machine it allows unique insights into microbial communities via sequencing of random (shotgun) DNA fragments taken directly from an environment.

Direct high-throughput sequencing of DNA has becoming cost effective ,with recent changes to a number of sequencing machines from companies like Roche (454) or Illumina (Solexa), data sets with several millions or even hundreds of million DNA fragment are becoming routine. With the appropriate analysis tools (CAMERA, IMG/M(1), MG-RAST(2), MEGAN(3)) shotgun metagenomics allows the study of microbial communities in ways that was never feasible before (4-7).

The process of democratization of sequencing however the requirements for the community of metagenomics analysis providers are steadily increasing (awkward). Novel computational challenges are arising in many locations at the same time. Even previously simple tasks like an all-vs-all comparison of DNA fragments within a metagenome are difficult at 150 million fragments per sample. Yet at the time of writing, the vendors are preparing to release yet another generation of their respective chemistry, thus pushing the limits even further.

The groups providing computational analysis need to face a number of challenges:

- establish a working solution for data exchange (note the work of the M5 platform in the Genomics Standards Consortium)
- the same is needed for the exchange of primary analysis (e.g. BLAST results), recomputing the entire analysis if you are adding the n+1 metagenome to an existing study is getting less and less feasible

At the same time the new data also provides a new set of entirely different challenges, only a subset of the vast amounts of DNA can be mapped to known proteins.

Finding solutions that allow systematic mining of large scale DNA data sets (metagenomes) for novel proteins (or fragments of DNA coding for novel proteins) is one of the tasks that will arise in the not too distant future for people analyzing metagenomes.

References:

1. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, et al. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* 2008;36(Database issue):D534-8. PMID: 2238950.
2. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics* [electronic resource]. 2008;9:386. PMID: 2563014.
3. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome research.* 2007;17(3):377-86. PMID: 1800929.
4. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, et al. Functional metagenomic profiling of nine biomes. *Nature.* 2008;452(7187):629-32.
5. Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Microbial ecology: human gut microbes associated with obesity. *Nature.* 2006;444(7122):1022-3.
6. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. *Nature.* 2009;457(7228):480-4.
7. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science.* 2004;304(5667):66-74.

Energy and electron flow in microbial communities: Understanding and controlling the many faces of corrosion

Ken Nealon
Wrigley Professor of Geobiology
University of Southern California

Abstract: Corrosion of materials of various kinds is estimated to cost the U.S. on the order of 275 billion dollars per year (FHWA report RD-01-156), a staggering yearly cost, nearly equaling the entire cost of reparation of hurricane Katrina. This doesn't count the cost of medical "corrosion" (bone and teeth loss), and the potential of global climate impact by "corrosion" of carbonate minerals. Understanding corrosion in its many forms, and especially being able to distinguish those forms that are biologically impacted from those that are purely physical/chemical is an absolute necessity and prerequisite to being able to deal with the problem in a preventative way. It is my strong opinion that only a concerted effort between experimental scientists, (microbiologists, chemists, geologists and engineers) coupled with a major effort in modeling and computing will provide the pathway for ameliorating this problem.

Background:

Electron transfer: Perhaps the thing that microbes (Bacteria and Archaea) do best is electron transfer. They have perfected the art of moving electrons – whether the energy is supplied by photons, by organic carbon, or by inorganic electron donors, bacteria seem to know all the tricks. In addition, about 20 years ago, it became clear that bacteria were capable of extracellular electron transport (EET), the movement of electrons to solid electron acceptors such as iron or manganese oxides, or even the anodes of microbial fuel cells (1). Since this first report, there have been literally hundreds of publications concerning EET, and it is now clear that microbes interact with surfaces of all kinds, not only as places to sit, but as sites of energy (electron donors), or sites of respiratory electron acceptors.

Corrosion: In the general sense, some types of corrosion are simply an expression of microbial metabolism – expression of a community that is almost universally a mixed species, mixed function group of microbes working together to move electrons around, harvesting energy and producing biomass. The byproducts of this metabolism can be electrons donated to minerals or materials, electrons extracted from minerals or materials, or protons (or organic acids) that might destabilize minerals or materials. What is not well known or understood is which types of corrosion are in fact catalyzed by microbes, and which are simply physical or chemical, and in the former, which microbes are involved (and what mechanisms are being employed).

Biofilms and microbial consortia: Of central importance with regard to corrosion is the almost universal interaction of different species under anaerobic conditions to accomplish the movement of electrons through the system. There are general statements, noting that the major organisms involved with the corrosion of steel are delta proteobacteria in the group

Desulfobvibrio, but in fact few studies have utilized modern molecular approaches to identify the microbes and or the metagenomic content of various types of corrosion. To my knowledge, there have been no attempts to establish such a data base, and use it to understand the nature of the corrosion process at a scale that might yield promising mitigation to this huge problem.

Logic: The logic used to define this problem is extremely simple: microbes are opportunists that will exploit nearly any type of chemical energy available on Earth, and given that corrosion reactions are universally energy yielding, it is to be expected that communities of microbes will have evolved to take advantage of these energy yielding reactions. The operating assumption then, is that the process is not a single-cell microbial process, but an integrated community “effort” that yields the result. Perhaps orders of magnitude more complex than the situation in a single cell, involving, settling, biofilm formation, cell to cell communication, and the energetic of communities rather than of single cells or even single species.

Question to be addressed: While my interest is focused on the area of corrosion, the more general question is one of the dynamics of energy flow (and carbon flow) through complex microbial consortia, and the accompanying impact of these reactions on materials of many kinds. The challenges are of a magnitude that it will really not be possible to do without the interaction of large scale computing and model establishment and testing. I am such a novice that I have no idea of the importance of whether it is 10, 100 or 1000 petflops!

Top 10 problems?: It is probably not an exaggeration to say that this is one of the premier interdisciplinary problems in the world today. A general understanding of energy flow in microbial ecosystems would impact not only corrosion (a \$300 billion problem), but medical science, and global carbon cycling as well: international cooperation participation is expected to be immediate and widespread.

How does petascale modeling and simulation help? Here is where I have a lot to learn. To my knowledge, the communities working in these areas have not had serious interactions. Given the complexity of the system(s), it would seem to be an ideal place to focus such efforts.

Other approaches: Traditionally, corrosion has been considered to be an engineering problem, and while MIC (microbially induced corrosion) is well known among corrosion engineers, very little in the way of mechanistic understanding is available. In fact, it is difficult to find extensive information on the microbiology of corrosion, and almost no molecular population analyses or metagenomics can be found. If a computational/experimental approach were established, it would stand to reason that the right kinds of data would appear, and some mechanistic understanding would result.

1. Myers CR & Nealson KH (1988) Bacterial Manganese Reduction and Growth with Manganese Oxide as the Sole Electron Acceptor. (Translated from Eng) *Science* 240(4857):1319-1321 (in Eng).

Building quantitative models of microbial ecosystems

Overview. One of the key challenges in microbial ecology is understanding how the numerous taxa act and interact to sustain a complex microbial ecosystem, even in well-mixed environments such as the ocean. Related to this central problem are the questions of how such communities assemble (deterministically or do founder-effects dominate?), and how communities change over time due to the predictable dynamics expected from biotic interactions or due to genetic mutation and acquisition of novel traits.

Peta-scale computing opens the door to building realistically-sized (though still multi-scale) simulations that could provide answers to some of these questions, but further empirical and theoretical studies are probably needed to ensure that these models accurately describe natural systems. More specifically, significant increases in computing power would, in principle, allow researchers to combine models of microbial metabolism, community structure, physical structure and fluid flows, and evolutionary dynamics, in order to make quantitative predictions about how species levels and metabolites should change over time. All of these models, however, are emerging areas of active research, and may not yet have the precision to be tied together in an “off-the-shelf” and independent fashion. An approach more likely to be successful is to use observed dynamics of microbial populations and metabolites to solve an inverse problem to estimate the biological interaction terms and the metabolic processes attributed to different groups.

A few of the main computational challenges are described below:

Genomics. High-throughput sequencing approaches have enabled characterization of microbial ecosystems to unprecedented detail at the DNA level. Single experiments can yield gigabases of DNA corresponding, for example, to tens of millions of individual marker genes describing the taxonomic makeup of a community. Later generations of DNA sequencers promise expanded throughput (the \$100 human genome), which could enable an era of microbial population genomics in which thousands or hundreds of thousands of microbial genomes are sequenced from each environment/timepoint. Performing simple preliminary analysis (genome assembly, gene prediction/annotation, homolog detection, ...) could prove difficult without peta-scale computing resources. More sophisticated analyses, including detection of natural selection or rates of recombination among strains or populations would raise the need for additional computational resources.

Population dynamics. A realistic model of population dynamics in a microbial ecosystem would combine physical models of fluid flow (on multiple spatial scales from cellular to global) with stochastic evolutionary dynamics including colonization, selection, and migration terms. In addition, recombination can generate new genotypes within populations, and modeling a large number of recombinant genotypes is important for basic questions in bacterial population genetics.

Metabolites. Complete genome sequences provide the raw material for understanding the metabolic role of bacteria in the environment, but new algorithms will be needed to automatically generate metabolic reconstructions sufficient for metabolic modeling (*e.g.*, new approaches to metabolic ‘hole-filling’). High-resolution cellular metabolic models based on complete genome sequences must be abstracted to a smaller list of consumed and excreted chemical species for inclusion in broader ecosystem models tied to physical models of the environment.

Eric Alm

Doherty Assistant Professor of Ocean Utilization, MIT

Martin Polz

Professor of Civil and Environmental Engineering, MIT