Perspective Article

Bioinformatics computation of metabolic models from sequenced genomes

Peter D. Karp {pkarp@ai.sri.com}
Bioinformatics Research Group, AI Center, SRI International, Menlo Park CA, USA

Christos A. Ouzounis {ouzounis@certh.gr}
Biological Computation & Process Laboratory (BCPL), Chemical Process Engineering Research Institute (CPERI), Center for Research & Technology Hellas (CERTH), Thessalonica, Greece

Abstract
In the early days of the human genome project (HGP), during the late 1980s and early 1990s, there was skepticism that the genome project would produce biologically meaningful information. The reality is that bioinformatics has allowed us to extract far more biology from sequenced genomes than any published predictions in the early 1990s. Thanks to the efforts of many researchers in several subfields of bioinformatics, we can now process a sequenced genome through a series of computations to produce a quantitative metabolic flux model. Thus, surprisingly, bioinformatics has achieved what might have been held up as a holy grail of the field, before the goal was even articulated.

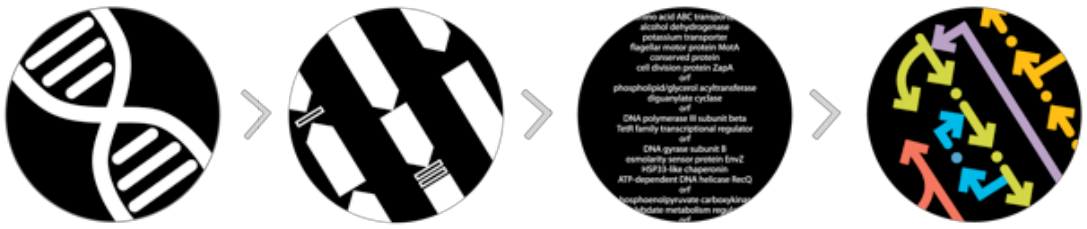
In the early days of the human genome project (HGP), during the late 1980s and early 1990s, there was skepticism that the genome project would produce biologically meaningful information (Cantor, 1990). Yes, the HGP would produce plenty of data, but how would we extract meaning from all those bytes? Back then, the sum of nucleotide sequences in GenBank was only 37 megabases (Watson, 1990). How would we convert billions upon billions of interleaved A, C, G, and T into biology? In 1990 it was suggested that the cost of finding genes in the human genome would be prohibitive because gene-finding would have to be performed experimentally (Weis, 1990), and even in a 1993 plan for the HGP, gene finding was listed in the experimental part of the plan, not the computational section (Collins and Galas, 1993). Finding genes within the genome and determining their functions were viewed as hard problems (Rowen, et al., 1997).

In parallel, another discipline was emerging: 'systems biology', or the study of biological systems by modeling the interactions among their components to infer emergent properties of the systems (Ideker, et al., 2001). A new scientific community crystallized around systems biology and catalyzed the convergence of the fields of sequencing and

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1501v1 | CC-BY 4.0 Open Access | rec: 12 Nov 2015, publ: 12 Nov 2015

simulation based on the realization that the genome sequence would provide the blueprint upon which an entire organism could be 'reconstructed' and that this complexity could be captured in computational models (Claverie, 2000; Karp, 2001).

The reality is that bioinformatics has allowed us to extract far more biology from sequenced genomes than any published predictions in the early 1990s. Thanks to the efforts of many researchers in several subfields of bioinformatics, we can now process a sequenced genome through a series of computations to produce a quantitative metabolic flux model. Although this process is imperfect – the resulting models contain errors and omissions outlined below – the models have significant predictive value. Computational inference of the metabolic network of an entire organism from its genome sequence was inconceivable during the instigation of large-scale sequencing and the HGP. Thus, surprisingly, bioinformatics has achieved what might have been held up as a holy grail of the field, before the goal was even articulated!



Computational inference of a metabolic model from a genome sequence	Gene finders compute the locations of genes within the genome	Protein function prediction and reactome inference compute the set of biochemical reactions catalyzed by the enzymes of the organism	Metabolic model generation fills gaps in the reactome and computes steady-state metabolic fluxes for the network

Figure 1: Steps from a Genome Sequence to a Metabolic Flux Model

Given the assembled genome sequence of an organism (Zerbino, et al., 2012), the steps required to convert that sequence to a metabolic flux model are as follows (**Figure 1**).

Step 1: Gene finding identifies the beginning and end of most genes within the genome. For an assembled genome sequence of any species, algorithms are available that detect translation start sites and internal features of genes from a combination of intrinsic properties of the sequence, plus experimental evidence such as expressed sequences (Haas, et al., 2008). This process is coupled to multiple iterative steps that validate initial predictions with additional information, for example comparative analysis (Yandell and Ence, 2012). The resulting sequence is typically fine-tuned to more accurately reflect eukaryotic intron boundaries and other complex features, such as untranslated regions (UTRs) (Yandell and Ence, 2012).

Step 2: Protein function prediction infers the biochemical activities of the protein products of genes identified in the preceding step on a genome-wide scale (Andrade, et al., 1999). The fundamental premise of function prediction for protein sequences

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1501v1 | CC-BY 4.0 Open Access | rec: 12 Nov 2015, publ: 12 Nov 2015

relies on the evolutionary connections within protein families. Bioinformatics methods infer the function of a protein by detecting the wider protein family to which the protein belongs. These connections are established through inexact string comparisons of the protein sequences, or by building computational models of the sequence patterns shared by multiple proteins within a family (Eddy, 2004). The methods infer that if the evolutionary distances between a query protein and a previously characterized protein or protein family are sufficiently small, they are likely to share a common function. Despite well-understood limitations, the process works well, especially for those protein families with high functional specificity (Schnoes, et al., 2009).

Step 3: Reactome inference (Karp, et al., 2011) uses the predicted protein functions to compute the reactome of the organism, the set of biochemical reactions catalyzed by the enzymes of the organism. Reactome inference translates protein functions to chemical reactions using associations stored within metabolic databases such as MetaCyc (Caspi, et al., 2012) and KEGG (Kanehisa, et al., 2012). Pathway prediction computes the metabolic pathways found in the reactome of the organism (Karp, et al., 2011), again with the aid of databases of known pathways such as MetaCyc and KEGG.

Step 4: Metabolic model generation produces a steady-state metabolic flux model (Durot, et al., 2009; Henry, et al., 2010; Latendresse, et al., 2012; Orth, et al., 2010). Because the model is at steady state, the concentrations of all metabolites are defined to be unchanging: thus, the fluxes of the reactions that produce each metabolite are balanced by the fluxes of the reactions that consume each metabolite. Those balance relations can be expressed as a set of constraint equations that are generated computationally from the reactome of the organism. The equations are submitted to a linear optimization package, along with the instruction to maximize the amount of biomass produced by the biosynthetic component of the metabolic network. The optimization package computes assignments of fluxes to each reaction in the metabolic network that optimizes the output of the network subject to the steady-state constraint.

An important component of the model generation step is gap filling (Latendresse, et al., 2012; Satish Kumar, et al., 2007). Genome-based metabolic network reconstructions are usually missing reactions because of incompleteness and inaccuracy in all of the preceding steps; a single missing reaction can prevent the network from producing one or more components of the organism's biomass. Gap-filling programs identify those missing reactions to produce a metabolic network that is sufficiently connected to produce a functional model (Agren, et al., 2013; Benedict, et al., 2014; Latendresse, et al., 2012).

Steady-state models do not have the large parameter requirements of kinetic models, which require hundreds or thousands of quantitative constants defining the properties of metabolic enzymes for accurate operation. Measuring those constants experimentally at a genome scale is currently intractable. In contrast, flux-balance analysis requires just four

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1501v1 | CC-BY 4.0 Open Access | rec: 12 Nov 2015, publ: 12 Nov 2015

inputs: the reaction network of the organism, a list of nutrient compounds for the organism, a list of compounds secreted by the organism, and a list of the compounds that compose the organism's biomass. Accurate recording of reactions and metabolites as well as their mapping and reconciliation to popular metabolic resources is also essential (Bernard, et al., 2014).

Applications of steady-state metabolic flux models range from anti-microbial drug discovery to metabolic engineering (Saha, et al., 2014). They can be used to predict the phenotypes of gene knock-out mutants by iteratively removing every reaction from the metabolic network, and determining whether all biomass components can still be produced (McCloskey, et al., 2013). One reason these models have limited accuracy is that they assume every metabolic reaction in the organism is potentially active at a given time, but that is not the case because regulatory processes shut down significant numbers of reactions under a given growth condition. Including regulation within these models is an active area of investigation (Machado and Herrgard, 2014). These models can also predict the ability of the organism to grow under different collections of substrates, growth rate and nutrient uptake rate of an organism. Automatically generated models were shown to predict microbial growth on various substrates, and gene essentiality, with an overall accuracy of 66% (Henry, et al., 2010).

The processing steps from genome to metabolic model form a computational pipeline, yet they represent more than a matter of programming. This achievement required significant conceptual advances. Computational gene finding resulted from the development of algorithms trained on species-specific statistical features of coding sequences, by gene detection based on gene-expression information, and by cross-species comparisons. Protein function prediction was enabled by the development of statistically rigorous sequence-matching algorithms, by the ability to computationally model protein families, and by large sequence databases. Reactome inference was enabled by the development of metabolic databases cataloging large numbers of reactions and enzymes. Metabolic model generation was enabled by the development of a steady-state modeling paradigm in which linear optimization is used to infer the flux distribution through a metabolic network; gap filling of reactions, nutrients, and secretions is used to automatically compute minimal-cost completions of models that would otherwise be incomplete and nonfunctional.

Note that development of new algorithms was not sufficient: databases play a key role in these (and many other) areas of bioinformatics. Bioinformatics problem-solving power comes from combining databases and algorithms: larger and more accurate sequence databases increase the power of sequence-comparison algorithms; larger and more accurate metabolic databases increase the power of reactome-inference algorithms.

To be sure, medical advances resulting from the human genome might have fallen short of expectations. And we must keep in mind that metabolic models describe only a fraction of the workings of a microbial cell, and an even smaller fraction of the machinery of multicellular organisms – such as developmental or signaling processes. Furthermore, steady-state models cannot readily describe dynamic behavior. Yet, the

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1501v1 | CC-BY 4.0 Open Access | rec: 12 Nov 2015, publ: 12 Nov 2015

general contribution of computation to interpreting genome data (Zerbino, et al., 2012) was underestimated at the dawn of the genomic era. Indeed, the costs saved by replacing experimental gene identification by computation in the human genome project probably paid for a significant fraction of all bioinformatics research. Despite its limitations, bioinformatics has more than fulfilled its promise. The interplay of bioinformatics with experimental molecular biology has turned into a rich dialogue (Ouzounis, 2012). Computation complements, suggests, and ideally obviates the need for experimentation.

Author Contributions

PDK conceived the article and authored the first draft; both authors contributed to extensive bibliographic searches and subsequent revisions of the article.

Parts of this work have been supported by the FP7 Collaborative Project MICROME (grant agreement #222886-2 to C.A.O.), funded by the European Commission.

References

Agren, R., *et al*. (2013) The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*, *PLoS Comput Biol*, **9**, e1002980.

Andrade, M.A., *et al*. (1999) Automated genome sequence analysis and annotation, *Bioinformatics*, **15**, 391-412.

Benedict, M.N., *et al*. (2014) Likelihood-based gene annotations for gap filling and quality assessment in genome-scale metabolic models, *PLoS Comput Biol*, **10**, e1003882.

Bernard, T., *et al*. (2014) Reconciliation of metabolites and biochemical reactions for metabolic networks, *Briefings in bioinformatics*, **15**, 123-135.

Cantor, C.R. (1990) Orchestrating the Human Genome Project, *Science*, **248**, 49-51.

Caspi, R., *et al*. (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases, *Nucleic Acids Res*, **40**, D742-753.

Claverie, J.M. (2000) From bioinformatics to computational biology, *Genome Res*., **10**, 1277-1279.

Collins, F. and Galas, D. (1993) A new five-year plan for the U.S. Human Genome Project, *Science*, **262**, 43-46.

Durot, M., Bourguignon, P.Y. and Schachter, V. (2009) Genome-scale models of bacterial metabolism: reconstruction and applications, *FEMS Microbiol. Rev*., **33**, 164-190.

Eddy, S.R. (2004) What is a hidden Markov model?, *Nat. Biotechnol*., **22**, 1315-1316.

Haas, B.J., *et al*. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments, *Genome biology*, **9**, R7.

Henry, C.S., *et al*. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models, *Nat. Biotechnol*., **28**, 977-982.

Ideker, T., Galitski, T. and Hood, L. (2001) A new approach to decoding life: systems biology, *Annual review of genomics and human genetics*, **2**, 343-372.

Kanehisa, M., *et al*. (2012) KEGG for integration and interpretation of large-scale molecular data sets, *Nucleic Acids Res*, **40**, D109-114.

Karp, P.D. (2001) Pathway databases: a case study in computational symbolic theories, *Science*, **293**, 2040-2044.

Karp, P.D., Latendresse, M. and Caspi, R. (2011) The pathway tools pathway prediction algorithm, *Standards in genomic sciences*, **5**, 424-429.

Latendresse, M., *et al*. (2012) Construction and completion of flux balance models from pathway databases, *Bioinformatics*, **28**, 388-396.

Machado, D. and Herrgard, M. (2014) Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism, *PLoS Comput Biol*, **10**, e1003580.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1501v1 | CC-BY 4.0 Open Access | rec: 12 Nov 2015, publ: 12 Nov 2015

McCloskey, D., Palsson, B.O. and Feist, A.M. (2013) Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*, *Mol. Syst. Biol.*, **9**, 661.

Orth, J.D., Thiele, I. and Palsson, B.O. (2010) What is flux balance analysis?, *Nat. Biotechnol.*, **28**, 245-248.

Ouzounis, C.A. (2012) Rise and demise of bioinformatics? Promise and progress, *PLoS Comput. Biol.*, **8**, e1002487.

Rowen, L., Mahairas, G. and Hood, L. (1997) Sequencing the human genome, *Science*, **278**, 605-607.

Saha, R., Chowdhury, A. and Maranas, C.D. (2014) Recent advances in the reconstruction of metabolic models and integration of omics data, *Curr. Opin. Biotechnol.*, **29C**, 39-45.

Satish Kumar, V., Dasika, M.S. and Maranas, C.D. (2007) Optimization based automated curation of metabolic reconstructions, *BMC Bioinformatics*, **8**, 212.

Schnoes, A.M., *et al*. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies, *PLoS Comput. Biol.*, **5**, e1000605.

Watson, J.D. (1990) The human genome project: past, present, and future, *Science*, **248**, 44-49.

Weis, J.H. (1990) Usefulness of the Human Genome Project, *Science*, **248**, 1595.

Yandell, M. and Ence, D. (2012) A beginner's guide to eukaryotic genome annotation, *Nat. Rev. Genet.*, **13**, 329-342.

Zerbino, D.R., Paten, B. and Haussler, D. (2012) Integrating genomes, *Science*, **336**, 179-182.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1501v1 | CC-BY 4.0 Open Access | rec: 12 Nov 2015, publ: 12 Nov 2015