AFRL-IF-RS-TR-2006-285
Final Technical Report
September 2006

DNA MEMORY AND INPUT/OUTPUT

Harvard University

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. M303

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

STINFO FINAL REPORT

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the Air Force Research Laboratory Rome Research Site Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-IF-RS-TR-2006-285 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/s/	/s/
THOMAS E. RENZ Work Unit Manager	JAMES A. COLLINS Deputy Chief, Advanced Computing Division Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
SEP 06	Final	Sep 01 – Dec 05

4. TITLE AND SUBTITLE
DNA MEMORY AND INPUT/OUTPUT

5a. CONTRACT NUMBER
F30602-01-2-0586

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
61101E

6. AUTHOR(S)
George Church

5d. PROJECT NUMBER
BIOC

5e. TASK NUMBER
M3

5f. WORK UNIT NUMBER
03

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Harvard University, Sponsored Programs
1350 Massachusetts Ave.
Cambridge Massachusetts 02138

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington Virginia 22203-1714

AFRL/IFTC
525 Brooks Road
Rome New York 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2006-285

12. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PA#06-639

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project started as a means to design and test various synthetic biology approaches to Biocomputing with DNA. Initial designs focused on biopolymer synthesis (DNA, RNA, protein) and in-vitro focused on computer aided design (e.g. CAD-PAM). Testing of various synthetic biology approaches integrated with improved functional-genomics quantization tools (e.g. MapQuant), which in turn integrated with systems-biology modeling (e.g. MOMA). The entire collection long-term should be capable of cycles of iteration. The metabolic modeling modules have developed as very useful in stand-alone software now in use in bioenergy applications and as an example of the successes and challenges on merging such software into the BioSPICE/BioCOMP community vision.

15. SUBJECT TERMS
DNA, Self Assembly, DNA Computing, Biocomputing, Systems Biology

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Thomas Renz
U	U	U	UL	85	19b. TELEPHONE NUMBER (*Include area code*)

Table of Contents

List of Figures

List of Tables

Acknowledgment

This work is undertaken by George Church (PI), John Aach (instructor), Peter Kharchenko & Alexander Wait (Harvard Biophysics graduate students), Jake Jaffe & Kyriacos Leptos (collaborating Harvard graduate students), Matthew Wright (MIT Chemistry graduate student), Xiaoxia Lin (postdoctoral fellow), Dan Segre (postdoctoral fellow), Jeremy Zucker (DFCI programmer), Aaron Brandes (programmer), Wayne Rindone (HMS Senior Staff), Sri Paladugu (KGI Graduate Student), Mike Hucka (KGI Assistant Professor), Tony Forster (Vanderbilt). We acknowledge Dr. Glenn Björk for help with the compilation of the tRNA modification data, the late Dr. James Ofengand, Dr. Stephen Blacklow. and Dr. William Studier for advice. We would like to thank Jay McPhee and Brent Martin for the maintenance of the clusters on which MapQuant was run. Moreover, we would like to thank professors Fritz Roth, Steve Buratowski and Steve Gygi for their advice and the latter for help with SEQUEST, Patrik D'Hasaeler and John Aach for thoughtful discussions and comments during the development of MapQuant, as well as Nikos Reppas. Finally, we would like to thank Jake Jaffe for sharing his expertise on the FT-ICR mass spectrometer.

1 Summary

The overall BioCOMP/BioSPICE project aims at integrating a broad set of "Systems Biology" measures and models ranging from functional genomics to simulation tools. Our Harvard sub-project focused on computer aided design (e.g. CAD-PAM) and testing of various synthetic biology approaches integrated with improved functional-genomics quantitation tools (e.g. MapQuant) in turn integrated with systems-biology modeling (e.g. Minimization of Metabolic Adjustment, MOMA). The entire collection long-term should be capable of cycles of iteration – CAD-Quant-Model -- CAD-Quant-Model … Initial designs focused on the major biopolymer synthesis pathways (DNA, RNA, protein), in vitro. That project was briefly expanded with supplementary funding to include experimental work which was then transitioned out to a Dept. of Energy grant which resulted in a Nature paper and commercial licensing (to CodonDevices). The metabolic modeling (MOMA) has developed as very useful in stand-alone software and as a prime example of the successes and challenges on merging such software into the BioSPICE community vision. During this project timeline we have developed ten major applications of these concepts and software tools:

a. A variety of chemical systems capable of replication and evolution, fed only by small molecule nutrients, is now designable and constructible. This could be achieved by stepwise integration of decades of work on the reconstitution of DNA, RNA and protein syntheses from pure components. Such an in vitro cell project (IVCP) would initially define the components sufficient for each subsystem, allow detailed kinetic analyses, and lead to improved in vitro methods for synthesis of biopolymers, therapeutics and biosensors. Completion would yield a functionally and structurally understood self-replicating biosystem. Safety concerns for synthetic life will be alleviated by extreme dependence on elaborate laboratory reagents and conditions for viability. The proposed minimal genomes are 113 kilobase pairs long and contain 151 genes.

b. Mathematical models of diffusion-constrained polymerase chain reactions provides the basis of high-throughput nucleic acid assays (licensed commercially to Agencourt – Beckmann - Coulter) and simple self-organizing systems (as in item 1 above)

c. The integration of the genomic and proteomic measurements and system-wide analyses was dramatically demonstrated in the first simultaneous determination of complete genomes and proteomes (initially of Mycoplasma mobile).

d. A cross-species expression data mining tool is now realized, and a first instance has been put in the public domain, called yMGV (yeast microarray global viewer). The comparative approach is applicable to filling of gaps in metabolic networks using

expression information. The metabolic expression placement (MEP) method relies on the co-expression to predict over 20% of all known Saccharomyces cerevisiae metabolic enzyme encoding genes within the top 50 out of 5594 candidates for their enzymatic function, and 70% of metabolic genes whose expression level has been significantly perturbed across the conditions of the expression dataset used.

e. Expression dynamics of a cellular metabolic network demonstrate predominance of local gene regulation. Metabolic genes display significant co-expression on distances smaller than the average network distance, a behavior supported by the distribution of transcription factor binding sites in the metabolic network and genome context associations. Positive gene co-expression decreases monotonically with distance in the network, while negative co-expression is strongest at intermediate network distances. Basic topological motifs of the metabolic network exhibit statistically significant differences in co-expression behavior.

f. While we and others have developed quantitation for full genome RNA since the mid 1990s, the protein equivalent has awaited software like MapQuant. This is Open-Source Software and has been exported to several groups including the largest genomics-proteomics center in the world, the Broad Institute, Cambridge where it is in routine use.

g. The important task of going from annotated genomes to metabolic flux models and kinetic parameter fitting has been addressed by a variety of software modules ranging from MOMA to ordinary differential equations (ODE) to comparative genomic and geometric constraints.

h. Accurate Multiplex Gene Synthesis from Programmable DNA Chips is now on-line. The CAD-PAM software for design of oligonucleotides for synthesis on chips and assembly into genomes is publicly available at:
http://arep.med.harvard.edu/cadpam.html.

i. Potential Bio-Security implications of some of the above work have been addressed by a novel, inexpensive, semi-automated means for surveillance of the synthetic DNA supply stream from chemicals, instruments, oligos and genes. http://arep.med.harvard.edu/SBP/Church_Biohazard04c.htm

2 Introduction

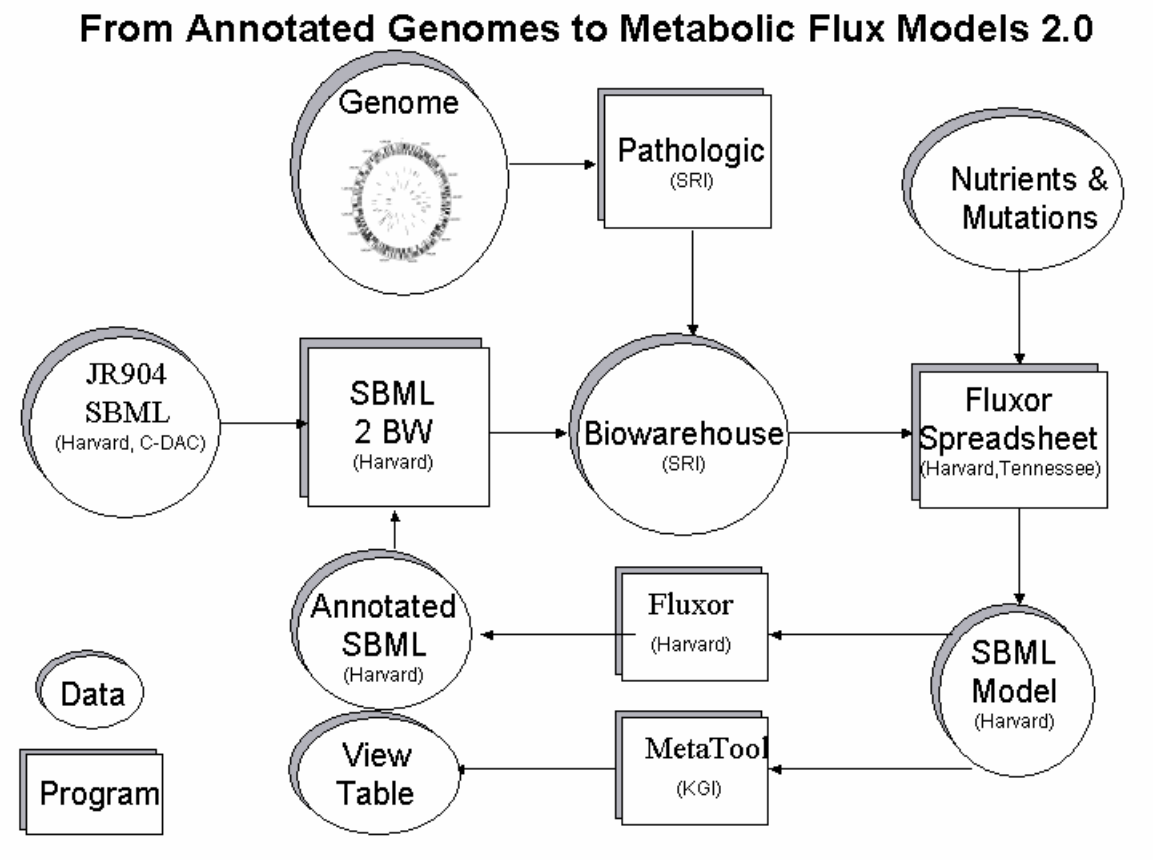
The approach was to develop a framework that accelerates the computational construction of models of genetic and metabolic processes that can be used to design, synthesize and optimize replicating in vivo and in vitro systems. These can be used for bioengineering applications that utilize biomolecules as information processing, bio-sensing, and structural components.

The above approach assumes some predicative power of computational models of genetic and metabolic processes. The design, synthesis and optimization of replicating in vivo and in vitro systems required new technology for fabrication of synthetic DNA, introduction into cells and high-throughput monitoring of the properties of the synthetic cells. This synthetic component empowers an important feedback loop to the computational modeling components.

The major focus of the Church Lab BioSPICE project has been the development of multisite use case software packages. Harvard and Dana Farber people have led a 5 site use case involving 16 investigators entitled “From Annotated Genomes to Metabolic Flux Models”. This use case starts with metabolic pathway networks that have been loaded into the Biowarehouse from SRI, either using the SRI tool called Pathologic to transform GenBank files into databases or in the future using a tool not yet developed to convert Systems Biology Markup Language (SBML) metabolic models into Biowarehouse databases. A tool was created that extracts metabolic networks as SBML files. These or any SBML files can be prepared for flux analysis and elementary flux mode prediction using an interactive spreadsheet based on the BioSpreadsheet developed at the University of Tennessee. The spreadsheet prepares annotated SBML that is used by the Metatool dashboard analyzer that was created by the Keck Graduate Institute to enumerate the elementary flux modes of the network and the Fluxor program developed at Harvard to make flux predictions for the model as well as predicting the fluxes for a knockout mutation requested using the spreadsheet using both the traditional linear optimization approach and the MOMA approach that minimizes the changes from the wild type flux predictions. The spreadsheet is used for a second time to display the flux predictions, and the standard BioSPICE Table View analyzer is used to display the pathways associated with the elementary flux modes.

Harvard has also collaborated with the Center for the Development of Advanced Computing in Pune, India, to create an SBML representation of the *E. coli* JR904 flux model. The Church Lab has prepared it using the spreadsheet and run a Fluxor analysis on it, but this first attempt did not yield realistic predictions. The spreadsheet is used to diagnose whether this is a problem with the Church lab model representation or with the Fluxor application, one expects to find fixes for both.

The use case distribution at http://arep.med.harvard.edu/moma/biospicefluxor.html is provided as a Linux tar file containing a dashboard analyzer for extracting reaction lists from the Biowarehouse as SBML files, the spreadsheet analyzer that prepares an SBML file for use by Fluxor or Metatool and displays the Fluxor predictions, and the Fluxor analyzer itself. It also contains the current state of the E coli JR904 SBML model, the tab-delimited listing of the model, and a tool that creates a new SBML file once the tab-delimited listing has been modified. The Biowarehouse and Pathway tools including Pathologic software can be downloaded from links at http://community.biospice.org/ or http://www.metacyc.org/. The tar file on arep.med.harvard.edu also includes instructions for setting up a MySQL database to hold the Biowarehouse Structured Query Language script that loads it with E coli pathway information derived from EcoCyc and a small hypothetical reaction system used as the Church lab demonstration case. Biowarehouse can also use an Oracle database, but the Church lab analyzer for extracting information from it as SBML was written using MySQL and there is no expectation it will work unchanged using Oracle. The Keck Institute has also set up a download page at http://public.kgi.edu/~spaladug/one.html, which includes windows installers for the Software Biology Workbench (SBW) including metatool and other agents and for BioSPICE dashboard analyzers that launch each of these tools. There are also Linux downloads for an SBW broker, the Metatool agent, and the NOM agent that handles providing an SBML file to Metatool. It is also necessary to copy the java jar files for metatool on the windows platform to the Linux platform to use these as dashboard analyzers. Once these SBW elements have been set up on a Linux machine they can be used in the same workflows as the Biowarehouse 2SBML Fluxor Spreadsheet, and Fluxor analyzers.



Project URL: http://arep.med.harvard.edu/darpabiocomp/
Quad Chart: http://arep.med.harvard.edu/darpabiocomp/Quad04_GC.ppt

Figure 1. From Annotated Genomes to Metabolic Flux Models

Objective

This project explores connections between computational systems biology and synthetic biology (bio-input/output, DNA memory & bio-manufacturing processes). This is done using the only class of programmable nanometer scale replicators (i.e. polymerase-ribosome-based). The major challenge is integration with silicon computing. The motivations are bio-monitoring of spatially patterned light, chemicals, and toxins. Software is aimed at BioCOMP/BioSPICE compatibility and emphasizes computational tools for analyzing complex metabolic networks and related synthetic biology goals.

Approach

The approach is to develop a framework that accelerates the computational construction of models of genetic and metabolic processes that can be used to design, synthesize and optimize replicating in vivo and in vitro systems. These can be used for bioengineering applications that utilize biomolecules as information processing, biosensing, and structural components.

The Systems Biology Markup Language (SBML) is a computer-readable format for representing models of biochemical reaction networks. SBML is applicable to metabolic networks, cell-signaling pathways, and genomic regulatory networks. The Systems Biology Workbench (SBW) is a modular, broker-based, message-passing framework for simplified communication between applications that aid in the above. The SBML module included the Network Object Model (NOM) and MetaToolSBW, a network analysis tool. These help to develop a pipeline from genome sequence and annotation to metabolic and genetic network optimization models. This employs linear and quadratic programming math modules. This is done in the context of experimental validation with an emphasis on integrating quantitative mass spectrometry, RNA tags, and effects of metabolic inhibition and related stresses.

In order to improve the performance of the fabrication and memory tools, in vitro replication/translation arrays will provide for experimental feedback. A 120kbp minigenome design shows capability for replication and protein-synthesis. This minigenome will be 6 times smaller than the smallest living cellular genomes with 1000-fold fewer molecular components. These in vitro systems are ideal for integrating with detailed computational models, due to simplicity, knowledge of the 3D structure of nearly all components and extreme experimental accessibility. Also coupling the extremes of modeling (from single base changes to 3D structures to molecular networks to population doubling selection) is likely to be dramatically more transparent and tractable.

Novel, Useful Applications & technology transfer: The focus is on practical applications that take advantage of the unique features of DNA and metabolic systems. Examples are: (a) proven Myr information archiving and retrieval; (b) interfacing with biochemical, photon, or thermal sensors. (c) A DNA recorder analogous to black-box flight recorder would take early advantage of the ability to record on DNA more easily than reading it. Only rarely would the archived materials be accessed.

In vitro minigenome synthesis:
The Church group led by Dr. Tian has shown that the multi-his-tag Western blots are a reliable assay. Hence, the minigenome genes have been moved into his-tagged "in vitro" linear-vectors. This included synthesis of tagged and untagged forms for all 23 genes of the 30S-ribosomal subunit. From these the Church group has synthesized all of the RNAs in vitro and most of the proteins. The low levels of protein synthesis observed for a few of these normally very abundant proteins is rapidly revealing key design criteria for codon usage and secondary structure of the mRNAs. The group has developed software for general gene and genome design tools that takes these observations into account.

A new method was developed for large scale synthesis of genes or genomes which has the potential of being 100-fold less expensive. This has been successfully tested by synthesizing two full-length genes from the minigenome (rs3 & rs5) from a mixture of 512 chemically synthesized 70mers. A report of invention has been filed with Harvard Medical School-Office of Technology Licensing. This is a major milestone for this project supplement. Joined by Hui Gong (for the oligo design), Nijing Sheng (gene

assembly expert, Research Asst. Prof. from the Univ. of Houston) and a Harvard undergraduate, and CS-graduate student. So this project is likely to continue to progress capturing this recent momentum.

Computational:
Church lab work on close-to-optimal networks as might occur in mutants, "Minimization of Metabolic Adjustment" (MoMA) has been extended to allow automated access to new genomes (Daniel Segre & Dennis Vitkup assisted by Jeremy Zucker, Tamar Mentzel, and Jeremy Katz) requiring only Kyoto Encyclopedia of Genes and Genomes or Genbank annotations as staring points.

Table 1. SBML models from BioCyc version 7.5 for 14 organisms

Agrobacterium-tumefaciens.xml
Bacillus-subtilis.xml
Caulobacter-crescentus.xml
Chlamydia-trachomatis.xml
Escherichia-coli.xml
Haemophilus-influenza.xml
Helicobacter-pylori.xml
Mycobacterium-tuberculosis-CDC1551.xml
Mycobacterium-tuberculosis-H37Rv.xml
Mycoplasma-pneumoniae.xml
Pseudomonas-aeruginosa.xml
Saccharomyces-cerevisiae.xml
Treponema-pallidum.xml
Vibrio-cholerae.xml

Plus, a program called biocyc2sbml.lisp which can take any organism in a Pathway/Genome database and generate the corresponding SBML model. The URL to download these models is http://genome.dfci.harvard.edu/~zucker/BPHYS/sbml.zip

An in vitro coupled replicating and translating system is based on pure bacterial *E.coli* translation. Novel developments include (1) a linear expression clone system compatible with the most powerful in vitro replication system, polymerase chain reaction (PCR), and (2) a modular method for computer gene design and automated gene synthesis including affinity-tagging for all ribosomal proteins.

The group merged Minimization of Metabolic Adjustment (MOMA) software with BioSPICE & SBML tools to allow optimization of metabolic network utilization in mutant genotypes and experimentally tested using metabolic fluxes (from Uwe Sauer's group) and a new high-throughput method for measuring growth rates of hundreds of mutants in parallel.

The group developed methods for 3D & 4D modeling of bacterial cells and replication translation of their circular chromosomes. In addition the Church lab has 1D to 4D

models of expansion of an in vitro DNA colony. High resolution atomic-force microscopy images have been obtained for heavy-atom labeled DNA samples.

The group completed integration of genome sequence for *Mycoplasma mobile* and *M. pneumoniae* with "complete" proteome comparisons. These are proving crucial for integration and 4D-modeling efforts and relevant to understanding these simple pathogens as biosystems.

3 Minimal Cell Design Tools

3.1 Methods, Assumptions, and Procedures

"How far can we push chemical self-assembly?"
This question was posed recently as one of the big 25 questions in science for the next 25 years (Service, 2005). Nowadays, big questions often are addressed by big experimental efforts. But before embarking on a big project, it is helpful to get specific. What push in chemical self-assembly might be most worthwhile and practical? Self-assembly in vitro of viruses and the ribosome, achieved decades ago, taught us some of the principles assumed to be used in general by cells (Lewin, 2004). For example, self-assembly occurs in a definite sequence and is generally energetically favored, obviating the need for enzymes and an energy source. Assembling some type of cell would seem to be the next major step, yet detailed plans have not been published. Here, we attempt to outline the synthesis of a minimal cell containing the core cellular replication machinery, review the pertinent literature, and highlight gaps in knowledge that need filling.

Utility
Synthesizing a minimal cell will advance knowledge of biological replication. Many hypotheses in replication and its subsystems can only be tested in such a synthetic biology project. The meaning of "synthetic" (from Greek *synthesis*, to put together) discussed here bypasses the current reliance of synthetic biology on cells or macromolecular cell products: the aim is to put together an organism from small molecules alone. The simplest approach for creating an artificial cell may be by evolving an RNA polymerase made exclusively of RNA (Szostak et al., 2001) to replace all protein components of in vitro replicating and evolving systems (*e.g.* to replace Qβ replicase (Mills et al., 1967)). But in comparison with a purified protein-based system, it is neither guaranteed to arrive sooner nor tell us more. A protein-based system will connect with, and reveal more about, existing biological systems. Life, like a machine, cannot be understood simply by studying it and its parts; it must also be put together from its parts. Along the way to synthesizing a cell, we might discover new biochemical functions essential for replication, unsuspected macromolecular modifications, or previously unrecognized patterns of coordinated expression.

How good a model would an artificial, protein-based, minimal cell be for natural cells? The only cellular alternative is a perturbed natural cell, an incredibly complex system even for the simplest of cells. A much simpler purified system based on a real cell would thus be easier to model and understand. It could certainly answer questions that cannot be answered in vivo or in crude extracts, such as which macromolecules and macromolecular modifications are sufficient for subsystem function. However, even the simplest minimal cell would still be highly complex, so its construction and study would be facilitated by substituting some of the necessary subsystems with simpler analogs. Should the simpler in vitro model turn out to be a poor model for the more complex in vivo system, one could always construct a more complex in vitro system that may better reflect in vivo.

Synthesizing a cell will also lead to new applications. Purified biochemical systems already offer major advantages, such as PCR and in vitro transcription. A better understanding and manipulation of all cellular replication subsystems (molecular biology's tool kit) should spin off new technologies. For example, in vitro genome replication may be useful for replicating very large segments of DNA with high fidelity. Combined in vitro transcription, RNA processing and RNA modification would allow preparation of rRNAs and tRNAs with defined modifications to test the roles of the modifications, and modified tRNAs to aid incorporation of unnatural amino acids into proteins. Purified translation systems have enabled reassignment of mRNA codons to encode unnatural amino acids by omission of competing natural amino acids (Forster et al., 2003); further improvements of the purified translation system could enable the genetic selection of protease-resistant, peptide-like ligands for drug discovery by pure translation display (Forster et al., 2004). The purified translation system may also facilitate expression of proteins difficult to express by standard approaches. Better control of lipid vesicle synthesis could advance liposome-based drug delivery. Since bacterial translation is the main target of antibiotics, greater understanding may assist development of new drugs to fight mounting antibiotic resistance. Ultimate success in cell synthesis could generate useful microorganisms, *e.g.* for renewable production of biodegradable plastics (Pohorille and Deamer, 2002).

3.1.1 Approach

The ideal approach for synthesizing a cell would allow all of the machine parts to be understood and tested. Like any engineering project, this requires detailed blueprints, raw synthetic capabilities and an overall diagnostic and debugging strategy. The use of entire genomes as the blueprints, some of which are small enough to synthesize *de novo*, is inconsistent with this approach. Self-replication of an unadulterated genome, however impressive, would not define the unnecessary genes, and the functions of about a third of the genes would remain unknown (Fraser et al., 1995; Jaffe et al., 2004). Building a machine from mysterious parts can only create a mysterious machine. What is needed is some way of defining a near-minimal genome and then a strategy that will lead inexorably to an understanding of all of its parts.

Theoretical and experimental studies have attempted to establish a minimal set of genes needed for a self-replicating system in a cushy constant environment of unlimited, small molecule nutrients (*e.g.* nucleotide triphosphates, amino acids, lipids and cofactors). Three basic approaches present themselves.

Comparative genomics searches for genes that have homologs in the genomes of groups of organisms. The approach estimates from 50 to 380 genes in a minimal genome (Jaffe et al., 2004; Koonin, 2000; Mushegian and Koonin, 1996; Tomita et al., 1999). It has the caveat that, among closely related genomes, some genes appear "required" for those species although they are not required for basic life. If one goes to longer evolutionary distances, many gene functions are replaced by non-homologous genes, hence making some essential genes look dispensable (*e.g.* the tRNA modification enzymes used by *Mycoplasma* are either different from *E. coli* or unidentified by sequence identity, but that doesn't mean the different ones are dispensable). An additional challenge is that about a third of the essential genes have unknown functions. It is thus expected that a minimal genome based on this approach alone would be unviable, and it would not be possible to identify the missing essential genes.

Genetics searches for essential genes by mutating one gene at a time. This approach estimates 330 genes in a minimal genome (out of *Mycoplasma genitalium's* total of 517; (Hutchison et al., 1999). Again, about a third of the essential genes have unknown functions. It is limited by false "essentials" due to the fraction of genes that were never mutated in the screen, due to creation of toxic partial complexes or pathways, and due to inadvertent effects on adjacent genes. The latter effects are prevalent in bacteria because a primary RNA transcript typically encodes multiple gene products. At the other extreme, false "dispensables" are disastrous when trying to assemble a viable minimal genome that lacks all of the individual "dispensables". For example, most RNA modification enzymes are individually dispensable, but simultaneous deletion of tens of them would be expected to be unsustainable due to cumulative reductions in efficiency or fidelity (a useful working definition of essentials for a minimal genome should encompass such lethal "dispensables"). Again, in using this approach alone, it would not be possible to identify the missing essential genes.

Biochemistry identifies from cell fractions those gene products essential for the reconstitution of biochemical reactions. It does not suffer from the above problems (except creation of toxic partial complexes), gives access to details of kinetic steps and allows debugging of isolated subsystems. However, the cellular subsystems must be integrated and thoroughly tested for accuracy on long templates before they can be considered physiological. Nevertheless, the biochemical approach has been successful at identifying macromolecules sufficient for reconstituting DNA, RNA and protein syntheses and, based on individual subtraction experiments, the components have either been shown to be necessary or could be so tested. Mindful of the remaining self-replication functions that need to be discovered (see below) it seems likely that a largely biochemical approach, now further empowered by mass spectrometry analyses and genetic and comparative genomic information, will be the most practical route to define a near-minimal, well-understood genome. We now review the relevance of current

knowledge and technology to this new minimal cell project (MCP), (Luisi, 2002).

A minimal genome

A MCP may be realized by reconstituting the macromolecular catalysts that synthesize DNA, RNA and protein. However, this overlooks the formation of the membrane compartment and the poorly-understood process in which it is divided by membrane proteins (Gitai, 2005), both of which are required for life. But lipids alone have been shown to be sufficient for formation of rudimentary membranous compartments capable of both transmembrane transport of small molecules and fission autocatalytically (Szostak et al., 2001), so membrane proteins may be dispensable. Polysaccharides should also be dispensable. If the simplest and best characterized examples of DNA, RNA and protein synthesis are selected, if translation of all codons is enabled for generalizability, and if efficiency and accuracy are not compromised, then this leads to the macromolecules and pathways of Figure A1.

A detailed list of the gene products in the hypothetical synthetic minimal cell of Figure A1 is shown in Table A1, left column. This list overlaps with a computational model of a minimal cell gene largely derived from a minimal organism, *Mycoplasma genitalium* (Tomita et al., 1999), but differs by omitting enzymes for synthesizing small molecules (*e.g.* lipids and glycolysis substrates) and by including DNA replication, RNA processing, RNA modification, extra tRNAs to decode the whole genetic code, some additional essential translation components, and chaperones. It should be emphasized that Table A1 is a working model only and that strict adherence will likely hamper progress. Examples of omitted, potentially stimulatory genes are given below. Conversely, examples of included, potentially dispensable genes may be gleaned by comparison with the streamlined *Mycoplasma* genome (Fraser et al., 1995).

Several conclusions can be drawn from the provisional list of genes selected for a minimal cell, most of which are attractive when contemplating a MCP. In genomic terms, the list is very short, containing only 151 genes and 113 kbp. All of the genes are derived from *E. coli* and its bacteriophages (except for the hammerhead RNA from a plant virus (Forster and Symons, 1987)), implying that the individual subsystems will be compatible. In contrast to lists derived by comparative genomics or genetic approaches, the biochemically-based list does not contain any genes of unknown function or challenging membrane proteins, so it is close to a fully understood, accurately replicating “platform” for life. The few known gaps constitute only about seven genes, all of which are predicted to be for RNA modification (Table A1, yellow in left column). From the viewpoint of structural biology, courtesy of recent breakthroughs in ribosome structure determination (Diaconu et al., 2005; Ogle and Ramakrishnan, 2005), significant three-dimensional information is lacking for only 3% of the products: a few RNA modification proteins and aminoacyl-tRNA synthetases (Table A1, yellow in right column). While some of the states and complexes remain to be solved at high resolution, a draft three-dimensional structure for any replicating system is a major milestone in the history of biology.

3.1.2 Tools

Genes for a MCP could be synthesized using either natural or unnatural gene sequences as starting points. Using natural gene sequences, genes can be readily synthesized by PCR, and large cloned operons of essential genes can be fused using synthetic linkers and homologous recombination. However, gene synthesis by cloning and PCR will soon be more expensive than raw synthesis from synthetic oligodeoxyribonucleotides (oligos). The latter also allows unnatural sequences, such as versions with altered codon bias to adjust mRNA secondary structures (Tian et al., 2004). Scalability and cost limitations of established methods for gene synthesis from synthetic oligos are now being overcome by oligo synthesis on chips followed by PCR amplification and error-correction (Carr et al., 2004; Richmond et al., 2004; Tian et al., 2004; Zhou et al., 2004).

3.2. Results and Discussion

3.2.1 Biochemical Subsystems

Several biochemical subsystems are required to synthesize a minimal cell, and they are reviewed here. For each subsystem, possible examples from natural systems will be compared, gaps in knowledge will be identified, and diagnostic and debugging strategies to fill the gaps will be suggested. Mindful of the goal of integration of the subsystems, emphasis is placed on subsystems that are homologous and that operate under standard physiological conditions.

Genome replication

In principle, the genetic material for a MCP could be either DNA or RNA. Although an RNA genome has the advantage of obviating genes for DNA replication, the challenges of preventing inhibitory double-stranded RNA structures and replicative mutations in artificial RNA genomes (Mills et al., 1967) are unsolved. So the genetic material for a MCP should be DNA.

A simple possible scheme for DNA replication that could be completely integrated with biological systems is shown in Figure A2. It shows rolling-circle DNA strand displacement (Zhong et al., 2001) initiated with RNA transcript primers synthesized in *situ* by an RNA polymerase. Processing of the resulting double-stranded DNA concatemers into monomeric DNA circles occurs by homologous recombination at Lox sites catalyzed by Cre recombinase (Sauer, 2002). This approach has advantages over existing rolling-circle (Dahl et al., 2004) or PCR (Mitra and Church, 1999) replication methods since it requires neither solid phase oligo synthesis nor changes in temperature, and is far simpler than natural DNA replication systems (Khan, 1997).

Rolling-circle DNA strand displacement could be engineered in a stepwise manner. First, a simpler version could be tested in which the T7 RNA polymerase and RNA processing are substituted by addition of short RNA primers. The efficiency of synthesis of monomeric DNA circles would be followed by gel electrophoresis (Dahl et al., 2004),

and replication fidelity at the base pair and whole genome levels should be tested with different polymerases. The biggest challenge anticipated is boosting the efficiency of monomeric circular template generation over byproducts, such as linear DNAs or oligomeric circles. Such defective byproducts would also be replicated and compete for nutrients (like PCR deletion products or defective interfering viruses). Defective byproducts potentially could be weeded out with appropriate selection schemes. For example, encapsulation of individual genomes within membranous cells would result in non-viability of cells containing deleted genomes.

Transcription

A single RNA polymerase should suffice for a MCP. Either *E. coli's* multi-subunit enzyme (Lewin, 2004) or the single polypeptide enzyme encoded by coliphage T7 (Studier et al., 1990) seem best, with the choice influenced by several considerations that also determine possible modes of regulation. In considering the whole transcription cycle for a minimal replicating system, the simpler, more predictable T7 RNA polymerase is arguably a better starting point than the *E. coli* RNA polymerase (a detailed comparison is provided in supplementary text).

RNA processing

A host of RNases cleave precursor RNAs in vivo (Li and Deutscher, 1996) with a complexity that could be reproduced in a MCP. However, inclusion of these RNases comes with the risks of cryptic cleavages, and a simpler approach may be easier to engineer (Figure A2, top). This approach generates all required unadulterated termini: tRNA 5’ and 3’ ends (Forster and Altman, 1990) and, if necessary, the 3’ end of a rRNA. The self-cleaving sequence (Forster and Symons, 1987) is included because precursor tRNAs with substantial 3’ extensions can be poor substrates for RNase P (Li and Deutscher, 1996) and RNA polymerase terminators are inefficient. The efficiency of RNA processing, monitored by gel electrophoresis, could be improved by trying several different precursor-specific sequences.

A minimal translatome

The most complex universal biological machinery is clearly translation. Translation-associated genes (the "translatome") account for a large fraction of cellular genes, 96% of the genes in Table A1, and all of the currently predicted gaps in knowledge for a MCP. The eukaryotic version is less attractive for engineering than the bacterial version because it contains some 30 initiation factor proteins and because eukaryotic ribosome assembly in vitro awaits the coordination of more than a hundred non-ribosomal macromolecules (Fromont-Racine et al., 2003). Of the bacterial systems, *Mycoplasma* has advantages over *E. coli* due its eight-fold-smaller minimal genome and its simple set of 29 tRNAs that is the only completely characterized set (Andachi et al., 1989)). Unfortunately, other important biochemical information for *Mycoplasma* is essentially unknown in areas where it is well-studied in *E. coli* (*e.g.* reconstitution of ribosomes and translation,

characterization and functional assays of rRNA modifications, characterization of RNA modification enzymes). Presently, this seems to favor the *E. coli* translatome for a MCP.

Purified translation

Efficient synthesis of proteins has been reconstituted from purified natural components (Kung et al., 1978) or recombinant His-tagged translation factors (Shimizu et al., 2005) from *E. coli*, but not yet from eukaryotes. The next steps with the *E. coli* system will be verifying accuracy by mass spectrometry and extending the short lifetime of the batch mode by continuous dialysis (Spirin et al., 1988). The versatility of the system will become apparent as more mRNAs are translated. If stronger mRNA secondary structures prove inhibitory despite the helicase activity of the ribosome (Takyar et al., 2005), introduction of an RNA helicase may be helpful. Given that aminoacyl-tRNA synthetases, translation factors, and ribosomal proteins are among the most abundant proteins in the cell, it will be important to verify that the purified system can produce high concentrations of all of these proteins.

An in vitro ribosome

The ribosome of choice is from *E. coli* because, in contrast with its eukaryotic cousins, it has been self-assembled from its purified components (Nierhaus and Dohme, 1974; Nomura and Erdmann, 1970; Traub and Nomura, 1968) and is homologous with the other components of the gene list (Table A1). Reconstituted ribosomes have only been assayed by synthesis of phenylalanine polymers from polyU templates (Lietzke and Nierhaus, 1988), so future assays need to test initiation and elongation at non-UUU codons, and also termination. Furthermore, the self-assembly protocol is finicky and non-physiological. In vitro assembly of the 30S subunit under physiological temperatures has been attained recently by adding the DnaK/DnaJ/GrpE chaperone system (Maki and Culver, 2005), although this system is dispensable in vivo (El Hage et al., 2001). Perhaps addition of natural polyamines might overcome the requirement for an unphysiologically high concentration of magnesium ions. All 54 of the ribosomal proteins have been cloned ((Culver and Noller, 1999; Semrad et al., 2004); the hypothesis that they (and other proteins in Table A1) can be synthesized in a purified translation system in active forms warrants testing.

rRNA production in a purified system is complicated by post-transcriptional nucleoside modifications. Since 5S rRNA lacks nucleoside modifications and is short, it is not surprising that it is active when transcribed in vitro (Zvereva et al., 1998). But the other two rRNAs are modified by about 20 enzymes in *E. coli*, half of which are unidentified. All 11 modifications of the *E. coli* small subunit 16S rRNA are dispensable for subunit assembly and aminoacyl-tRNA binding (Krzyzosiak et al., 1987). However, *E. coli* 23S rRNA lacking its 23 modifications is 30-fold less active than the natural version in N-Ac-Met-puromycin synthesis (Semrad and Green, 2002) due to one to six modifications in a relatively small RNA domain (Green and Noller, 1996). The enzymes that catalyze these six modifications are therefore included in Table A1, although the two known ones are individually dispensable (Del Campo et al., 2001). Other bacteria should also be entertained for a MCP, as these six *E. coli* modifications are not conserved and the unmodified 23S RNAs from two other eubacteria are quite active (Green and Noller, 1999; Khaitovich et al., 1999).

In vitro tRNAs

Which of the myriad tRNA genes and tRNA modification enzymes are likely to be sufficient to decode all 61 sense codons in a MCP? There are some 85 tRNA genes in *E. coli* coding for some 45 different tRNAs each bearing post-transcriptional modifications on about 10% of their nucleosides, and a fifth of the tRNAs still remain to be characterized at the modification level. At least 27 different types of nucleoside modifications are present in *E. coli* (Bjork, 1995). There are an estimated 40-50 tRNA modification enzymes in *E. coli*, about half of which remain to be identified. To make matters worse (or more interesting) for a MCP, the roles of the tRNA modifications are controversial.

Arguments for choosing essential tRNA modification activities are highly speculative. As few as 33 *E. coli* tRNAs may be sufficient to translate the entire genetic code accurately (Table A1, left; Figure A3). *E. coli* tRNAs could be substituted with the completely characterized set from *Mycoplasma capricolum*, which contains only 14 types of nucleoside modifications (Andachi et al., 1989), some of which differ from *E. coli*. However, the predicted savings in number of essential tRNAs and modification enzymes are minor (Table A1, middle column), and full compatibility with the heterologous *E. coli* translation apparatus seems unlikely (*e.g.* the codon UGA in *Mycoplasma* encodes Trp, not stop).

Each in vitro-synthesized nascent tRNA transcript should be modified with different combinations of modification enzymes and tested for efficiency and accuracy of codon recognition in translation, initially in a simplified purified translation system (Forster et al., 2001). Identification of the unknown modification enzymes is being hastened by bioinformatic and genomic approaches (Soma et al., 2003). It is also conceivable, though unlikely, that unknown small molecules would need to be identified biochemically for RNA modification (or other reactions). The remaining *E. coli* tRNA modification enzymes not listed in Table A1 might be predicted to be dispensable based on available data (Bjork, 1995; Giege et al., 1998). But given the uncertainties, it may be faster to get to a working near-minimal cell by using every known *E. coli* modification enzyme.

Post-translation

A MCP must promote correct protein folding and any necessary post-translational amino acid modifications. Early versions of a purified replicating system will contain cell-derived macromolecules, so establishing that such systems can be completely weaned from cells will require enough rounds of replication for “infinite” dilution of the starting macromolecules. This will test for dependence on folding by chaperones and on post-translational modifications. It is unclear which, if any, chaperones will be necessary, but GroEL/ES (El Hage et al., 2001; Kerner et al., 2005) are likely candidates (Table A1). The only known examples of required post-translational modifications for the proteins in Table A1 are the recently discovered methylations of translation release factors 1 and 2 catalyzed by release factor Gln methylase (Table A1) (Heurgue-Hamard et al., 2002; Nakahigashi et al., 2002). Other possibilities include ribosomal protein acetylations. Mass spectral comparisons between proteins made in the purified system and those made

in vivo will expose modifications and also assess fidelity, while the inactivity of a protein of expected mass would suggest a protein-folding deficit and the need for an additional chaperone. Any necessary missing components could be identified biochemically by mixing with fractionated crude extracts or through genetics.

Compartments and division

Membranes would allow evolution without serial transfers and purifications, extension of the system to new environments, and better modeling of cells. On the other hand, membranous boundaries are unnecessary for directed evolution (Mills et al., 1967) or, in theory, self-replication. Membranes also restrict applications (*e.g.* delivery of unnatural amino acyl-tRNAs, selection schemes based on binding and spatial arraying for nanofabrication). Addition to self-replicating macromolecules of lipids alone may be sufficient for encapsulation of the macromolecules within bilayer membrane vesicles, synthetic cell division and transmembranous small molecule transport (Szostak et al., 2001). The choice of lipids is wide open, but one should not underestimate the challenges involved in working with them (Luisi, 2002) nor the advantages in regulation to be gained by adding membrane-modeling proteins (*e.g.* pores, transporters and the yet-to-be-discovered complement of cell division proteins (Gitai, 2005)).

Integrating the subsystems

How might all of the biochemical subsystems in Figure A1 be combined to generate a self-sustaining system? This is clearly a new level of complexity in comparison with prior self-assembly projects. None of the subsystems described above are completed, yet their selection is based on a reasonable plan for their ultimate integration. The approach again would be stepwise, and there are many possible pathways that could be integrated in parallel (Figure A1). For example, transcription by T7 RNA polymerase couples well with a purified *E. coli* translation system (Shimizu et al., 2005). Theoretical integration of DNA synthesis, RNA synthesis and RNA processing was discussed above (Figure A2). These four different subsystems could then be combined to synthesize part of a fifth system (the ribosome) by synthesis of an antibiotic-resistant 16S rRNA and His-tagged versions of all 21 small subunit ribosomal proteins (Tian et al., 2004). The products of these integrated subsystems could then be assayed for correct in vitro reconstitution of small ribosomal subunits by (i) selecting for resistance of protein synthesis to the antibiotic, and (ii) detecting the presence of tagged proteins in purified small ribosomal subunits by Western blot with anti-His antibodies. As another example, rudimentary vesicles encapsulating replicating systems (*e.g.* Q replicase) were shown to be capable of multiplication (Luisi, 2002).

Numerous fine-tuning strategies can be envisioned. Relative strengths of DNA promoters and mRNA ribosome-binding sites for different genes could be modeled on the in vivo strengths, with necessary adjustments of synthetic rates (and thus concentrations of products) achieved by mutations in the binding sites. Additional modules might be useful, such as catabolism (nucleases and proteases), active conversion or removal of waste products (*e.g.* by energy regenerating enzymes or membrane transporters) and regulatory feedback (*e.g.* excess transcription -> excess T7 lysozyme mRNA -> excess lysozyme -> lysozyme binding to and inhibition of T7 RNA polymerase). Control of

macromolecular concentrations will be aided by in *silico* modeling and design (Tomita et al., 1999). Given that the subsystems discussed above were selected with integration in mind by choosing physiological reaction conditions and homologous components, and given that additional subsystems could always be borrowed from living cells as needed (*e.g. E. coli* RNA polymerase and regulatory modules such as riboswitches (Isaacs et al., 2004)), it seems likely that this approach will eventually produce synthetic self-replication and ultimately a self-sustaining minimal cell.

It is important to note that a minimal cell would be intentionally fragile. For example, the vesicle would be easily lysed and the small molecule feeding mix would be highly specialized indeed (including unstable cofactors such as N-5,10-methenyltetrahydrofolate and S-adenosylmethionine). These built-in safety features will prevent a minimal cell from replicating outside the laboratory. However, some or all of the synthetic genes for a MCP would be intentionally passaged through living cells for construction of recombinant DNA clones and for amplification. Constantly upgraded ethical and safety regulations in place for existing biohazards would also encompass this research (Cho et al., 1999); http://arep.med.harvard.edu/SBP/Church_Biohazard04c.htm .

3.2.2 Completion

In conclusion, a stepwise biochemical approach lends itself to the eventual identification of any remaining functions essential for the synthesis of a minimal cell sustained solely by small molecules. Five states of completion present themselves as tractable goals of a MCP. Namely, the identification of:
(1) the genes listed as missing in Table A1,
(2) any additional genes and organization necessary experimentally for minimal cell synthesis,
(3) any dispensable genes,
(4) biochemical parameters and computational models sufficiently detailed to predict the effects of alterations, and
(5) the missing three-dimensional structures of the gene products and their relevant complexes.

It is difficult to predict how long it will take to debug each of the individual biochemical subsystems or to put them all together, so it is important to bear in mind that there are short-term goals. Intermediate assembly steps could also be pursued while the gaps in RNA modification knowledge (Table A1) are being filled. For example, the project to assemble a ribosome under physiological conditions could be carried out without the missing 23S rRNA modification enzymes (Table A1) by substituting in natural 23S rRNA. Similarly, assembly of self-replication in the absence of functional in vitro-synthesized tRNA substrates could be carried out using cellular total tRNA to enable self-replication from substrates (rather than just small molecules) as a major step towards understanding biological self-replication. This would also allow directed evolution of all of the components except the tRNAs in a more flexible manner than is possible in vivo (*e.g.* for selecting ribosome mutants that incorporate unnatural amino acids more efficiently).

The biochemical subsystems necessary for a MCP are central, old fields that have lost impetus. Completion within a decade will only be possible through a coordinated filling of the key gaps in knowledge by the cutting-edge laboratories scattered around the world in these fields. It will also require stimulation of rate-limiting fields. For example, though rRNAs and tRNAs can constitute more than 70% of the dry weight of a cell, half of the estimated 60-70 RNA modification enzymes of *E. coli* and one fifth of the tRNAs remain to be characterized, despite the recent completion of about 300 bacterial whole genome sequences. The momentum of genomics and consequent deluge of computed hypotheses cries out for comparable breakthroughs in experimental tests. Synthetic systems biology projects such as a MCP promise such tests with the added bonus of new applications.

Escherichia coli	*Mycoplasma*	3D structure
Coliphage f29 DNA polymerase	+	+
Coliphage P1 Cre recombinase	-	+
>Coliphage Lox/Cre recombinase site	-	+
Coliphage T7 RNA polymerase	analog	+
>Coliphage T7 RNA polymerase initiation site	analog	+
>Coliphage T7 RNA polymerase class II termination site	analog	+
Lucerne viral hammerhead RNA	-	+
RNase P RNA	+	+
RNase P protein	+	+
>RNase P site/RNA primer for DNA polymerase	+	+
Small subunit 16S ribosomal RNA	+	+
All 21 small subunit ribosomal proteins (1-21)	+ except 1,21	+
Large subunit 5S ribosomal RNA	+	+
Large subunit 23S ribosomal RNA	+	+
Large subunit 23S rRNA G2445>m2G methylase: unidentified	unknown	-
Large subunit 23S rRNA U2449>dihydroU synthetase: unidentified	unknown	-
Large subunit 23S rRNA U2457>pseudoU synthetase	unknown	-
Large subunit 23S rRNA C2498>Cm methylase: unidentified	unknown	-
Large subunit 23S rRNA A2503>m2A methylase: unidentified	unknown	-
Large subunit 23S rRNA U2504>pseudoU synthetase	unknown	-
All 33 large subunit ribosomal proteins (1-7,9-11,13-25,27-36)	+ except 25, 30	+
Translational initiation factor 1	+	+
Translational initiation factor 2	+	+
Translational initiation factor 3	+	+
Translational elongation factor Tu	+	+
Translational elongation factor Ts	+	+
Translational elongation factor G	+	+
Translational release factor 1	+	+
Translational release factor 2	-	+
Translational release factor Gln methylase	+	+
Translational release factor 3	-	+
Ribosome recycling factor	+	+
33/45 Transfer RNAs (see Fig. 2)	Set of **29**	+
tRNA C34>lysidine synthetase	unidentified	+
tRNA A34>I deaminase	unidentified	+
tRNA U34>cmo5U (=V) synthetases: unidentified	-	-
tRNA U34>2sU Cys desulfurase	-	+
tRNA U34>2sU synthetase	unidentified	+
tRNA U34>cmnm5U GTPase	unidentified	+
tRNA U34>cmnm5U synthetase	unidentified	+
tRNA cmnm5U34>nm5U>mnm5U synthetase	unidentified	-
tRNA G37 N1-methylase	+	+
tRNA A37>t6A N6-threonylcarbamoyl-A synthetase: unidentified	unidentified	-
tRNA A37>i6A synthetase	-	+
tRNA i6A37>s2i6A>ms2i6A synthetase	-	+
All 22 aminoacyl-tRNA synthetase subunits (20 enzymes)	+ except Gly sub., Gln	+ except Gly sub., Ala
Met-tRNA formyltransferase	+	+
Chaperonin GroEL	+	+
Chaperonin GroES	+	+
151 genes = 38 RNAs + 113 proteins		

Table A1. Biochemically-derived list of genes that may encode a useful, near-minimal, self-replicating system dependent only on small molecule nutrients.

Gaps in knowledge are in yellow. **Left column:** chosen gene products and DNA sites. **Middle column:** relationship to the minimal genome of *Mycoplasma genitalium*; clear sequence homolog = "+"; known enzyme product without an evident sequence homolog = "unidentified"; no functional homolog = "-". **Right column:** high-resolution, three-dimensional, structural information; >25% of the structure solved = "+", <25% = "-".

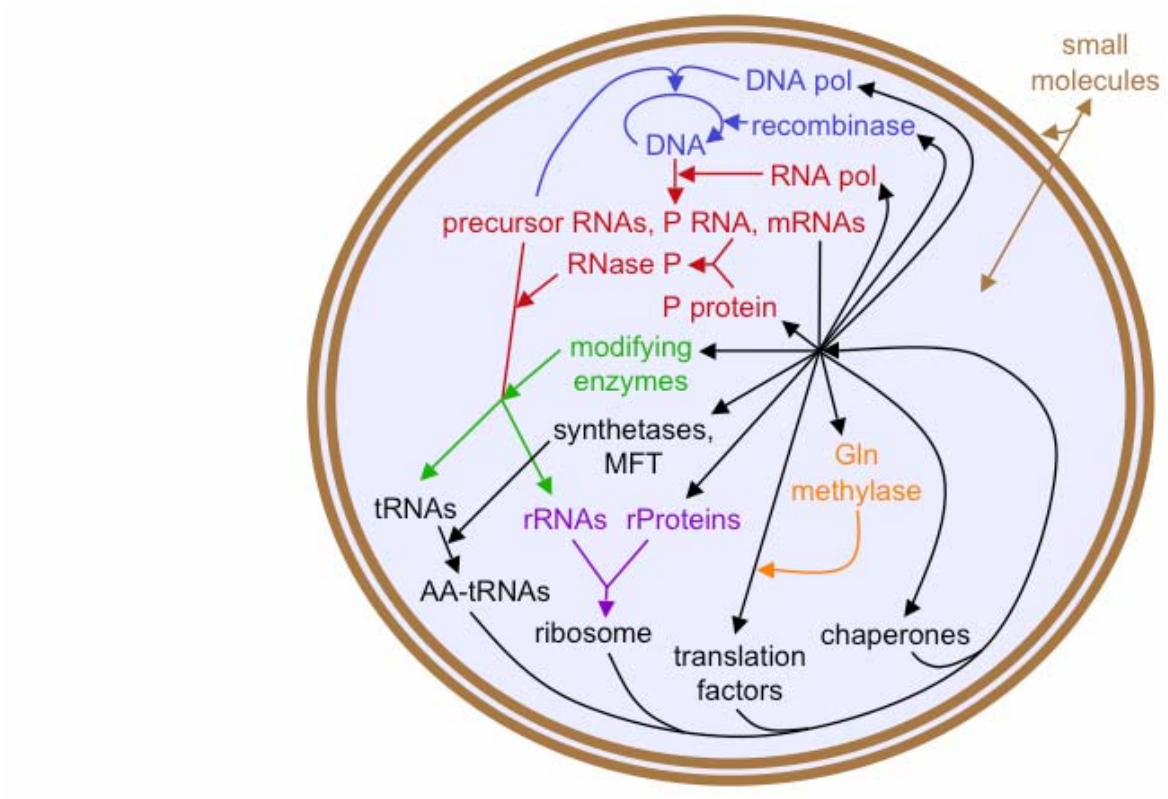


Figure A1. A minimal cell containing biological macromolecules and pathways proposed to be necessary and sufficient for replication from small molecule nutrients.

The macromolecules are all nucleic acid and protein polymers and are encapsulated within a bilayer lipid vesicle. The small molecules (brown) diffuse across the bilayer. The macromolecules are ordered according to the pathways in which they are synthesized and act. They are colored by biochemical subsystem as follows: blue = DNA synthesis, red = RNA synthesis and cleavage, green = RNA modification, purple = ribosome assembly, orange = post-translational modification, and black = protein synthesis. MFT = methionyl-tRNA$^{fMet}_{i}$ formyltransferase.

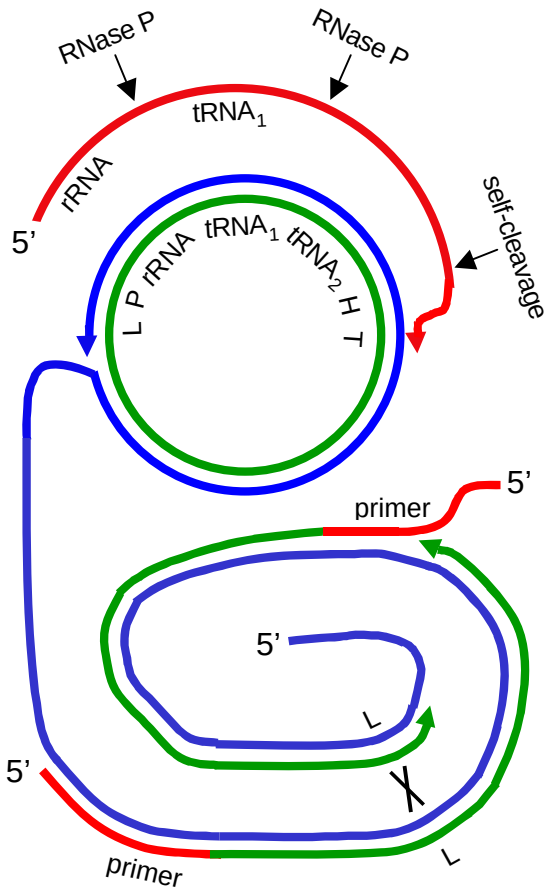


Figure A2. A generalizable, physiologically-compatible theoretical schema for accurate DNA replication and RNA synthesis in vitro.

Polymerase movements are illustrated by colored arrow heads. **DNA synthesis.** A nicked double-stranded DNA circle (middle) undergoes rolling-circle DNA synthesis by coliphage phi29 DNA polymerase (Dahl et al., 2004) to give an oligomeric single-stranded DNA (bottom, blue). RNA primers (red) then hybridize at two sites to prime lagging strand DNA synthesis (bottom, green). When two Lox sites (bottom, L) are completed, recombination occurs between them catalyzed by coliphage P1 Cre recombinase (black cross) to form a duplicate of the original circular template. **RNA synthesis.** The circular genetic operon (middle) contains a promoter for T7 RNA polymerase (P), a ribosomal RNA (rRNA) gene, two transfer RNA (tRNA) sequences, a self-cleaving hammerhead sequence (H), and a T7 terminator (T). RNA synthesis from P generates a precursor RNA (top, red) containing three cleavage sites (thin black arrows). The second tRNA sequence merely serves as a recognition site for RNase P cleavage. Cleavages yield the mature rRNA and $tRNA_1$. Any cleavage product containing a 3' hydroxyl group or primary RNA transcript can serve as a primer for DNA synthesis (bottom, red).

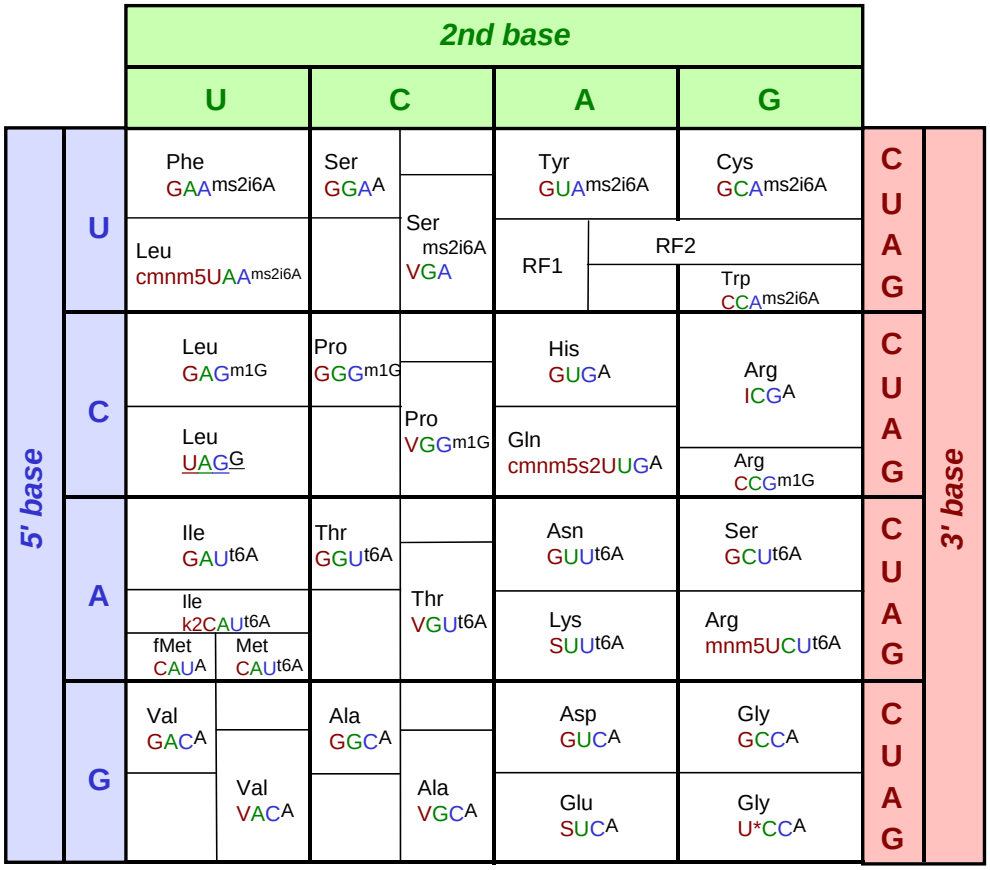


Figure A3. All nucleoside modifications of all 33 synthetic tRNAs that may be sufficient for accurate translation.

Outside (shaded): mRNA codons of the genetic code are illustrated in the standard format, except that the 3' U and C are switched to simplify depiction of decoding. **Inside:** tRNA nucleotides 34-37 (from 5' to 3') and their cognate amino acids. Nucleotides 34-36 are the anticodons, and nucleotides 37 are represented by black superscripts. Codon and anticodon positions that base pair with each other are colored similarly. Stop codon specificities of release factor (RF) proteins are included. The portions of the tRNA sequences not shown in the figure are unmodified. Expected modifications of in vitro transcripts by the enzymes in Table A1, and expected amino acid and codon specificities are given. * = unspecified modification, _ = unknown modification status, ms^2i^6A = 2-methylthio-N^6-isopentenyladenosine, m^1G = 1-methylguanosine, t6A = N^6-threonylcarbamoyladenosine, $cmnm^5U$ = 5-carboxymethylaminomethyluridine, V = cmo^5U = uridine 5-oxyacetic acid, I = inosine, $cmnm^5s^2U$ = 5-carboxymethylaminomethyl-2-thiouridine, k2C = lysidine, S = mnm^5s^2U = 5-methylaminomethyl-2-thiouridine, mnm5U = 5-methylaminomethyluridine.

Transcription by RNA polymerase from *E. coli* versus coliphage T7

Initiation

All *E. coli* genes are transcribed by *E. coli* RNA polymerase, given the appropriate sigma subunit (Lewin, 2004). For a MCP with the target genes in Table A1, use of natural promoters with the rpoD sigma subunit would probably maintain the natural relative initiation rates, although this needs to be confirmed. Any promoters that didn't function (perhaps due to the lack of an uncharacterized factor) could then be "fixed" by substitution with working promoters. Natural regulatory proteins could also be added to provide control mechanisms. However, cryptic initiation sites are anticipated for *E. coli* RNA polymerase but not for the more selective T7 RNA polymerase. The latter polymerase could be regulated by the binding of T7 lysozyme and by promoter binding by lac or other repressors (Studier et al., 1990). Relative initiation rates for T7 RNA polymerase could be set using T7 promoters of different strengths, although such promoters dictate the first few bases of the transcripts (usually pppGGG...). Nevertheless, this restriction in 5'-terminal sequence does not appear to be problematic for the synthesis of any of the RNAs needed for a MCP (see below).

Elongation and termination

T7 RNA polymerase transcribes about five times faster than the *E. coli* one, altering the relative *E. coli* rates of coupled transcription and translation, although this works well for phage T7 infections and when over-expressing genes in vivo. Termination by both enzymes is inefficient, so tandem terminators should be tried. T7 has the important advantage of high processivity through essentially any sequence until it reaches its natural terminator (class I or II (Lyakhov et al., 1998)). The *E. coli* polymerase terminates prematurely within genes containing an anti-termination signal (*e.g.* found in rRNA genes) if the anti-terminator factor(s) fail to act. Since the number of different anti-termination mechanisms in *E. coli* is unknown, the extent of anti-termination is also unknown. Though premature termination could be detected easily in MCP experiments and might be overcome in some cases by omission of transcription termination factors, other cases would require altering codon bias with the hope that the natural premature termination signal (sometimes an unknown sequence) is destroyed without otherwise affecting transcription or translation.

The cycle

The transcription cycle of *E. coli* is more complex than T7's because it requires association and dissociation of the sigma factor.

tRNA modifications

The roles of the tRNA modifications are controversial. The genetic approach almost always finds a particular tRNA modification enzyme to be dispensable (Bjork, 1995), and even the enzyme that synthesizes the universal U to T modification at position 54 is only essential due to a function separable from tRNA modification (Persson et al., 1992). The biochemical approach finds unmodified tRNAs to be active in translation systems (Cornish et al., 1995; Harrington et al., 1993), although careful comparisons between

individual unmodified tRNA transcripts and their modified counterparts (either purified natural isoacceptors or chemically-synthesized tRNAs (Wang, 1984)) are limited. Thus, the hypothesis that the modifications are unimportant is widely held. But the source of tRNAs for all in vitro protein syntheses are cellular total tRNAs (because so many different tRNAs are required), so attempting protein synthesis with in vitro-synthesized tRNAs will be helpful in testing this hypothesis. The contrary view, that several tRNA modifications will be of key importance for a MCP, seems most likely. The comparative approach argues for their essentiality (Bjork, 1995). Genetics rarely rules out pseudorevertants (suppression by secondary mutations (Gutgsell et al., 2001)) and has recently identified a few essentials (Bjork et al., 2001; Soma et al., 2003; Wolf et al., 2002) (not to mention the potential for pairwise lethals). Biochemical assay interpretation should be tempered by the presence in crude translation extracts of endogenous modification activities (Samuelsson et al., 1988) and the paucity of assays performed in pure charging and translation systems (Harrington et al., 1993).

What modification enzymes can be predicted to be essential for self-replication? Nucleotide 34 in the anticodon wobble position, and nucleotide 37 directly 3' of the anticodon, contain the most complex (hyper-) modifications, and all of the most likely essentials. Charging by aminoacyl-tRNA synthetases in *E. coli* requires mnm5s2U34 in tRNA-Lys and tRNA-Glu (and perhaps the related modification in tRNA-Gln), t6A37 in tRNA-Ile1, and lysidine34 in tRNA-Ile2 (Bjork, 1995; Giege et al., 1998). The latter is the only known example of a modification acting as an anti-determinant by preventing mischarging (with Met (Giege et al., 1998)), but a systematic search for others is needed by in vitro charging of unmodified tRNAs in a purified system containing all 20 aminoacyl-tRNA synthetases. Accurate wobbling during codon recognition requires lysidine34, mnm5s2U34 and its variants (mnm5U34 and cmnm5U34), cmo5U34 and inosine34 (Curran, 1998; Yokoyama and Nishimura, 1995). The active anticodon loop confirmation of tRNA-Lys is stabilized by direct interaction of mnm5s2U34 and t6A37(Sundaram et al., 2000). m1G37 is essential to prevent frameshifting (Bjork et al., 2001). t6A37 and msi6A37 (and its variant i6A37) stabilize A-U and U-A base pairs at the N36 position of codon-anticodon duplexes by stacking, presumably important for increasing translational efficiency at these codons (Grosjean et al., 1998).

4 Quantitative Proteomic Software

4.1. Methods, Assumptions, and Procedures

Whole-cell protein quantitation using mass spectrometry has proven to be much more challenging than mRNA quantitation. The detection efficiency varies significantly from peptide to peptide; the molecular identities are not evident a priori and are dispersed unevenly throughout the multidimensional data space. In this study we have developed open-source software, called MapQuant, which quantitates all organic species in large mass spectrometry datasets

We tested MapQuant on Bovine Serum Albumin, BSA samples in 21 Liquid Chromatography / Mass Spectrometry, LC/MS experiments, in triplicate at seven different concentrations (7 – 5000 fmoles) for quantitation purposes (labeled q-experiments) and two LC/MS experiments for identification purposes (labeled s-experiments). For each q-experiment, MapQuant generated a two-dimensional map of scans against m/z bins. Analysis entailed applying algorithms for noise filtering, watershed segmentation, peak finding and fitting, peak clustering and isotopic-cluster deconvolution and fitting using binomially distributed clusters of gaussioid peaks. MS/MS spectra were interpreted using the program SEQUEST.

Out of the 190 tryptic peptides used for the quantitation, 172 were identified by SEQUEST. Although the data were acquired on a low resolution spectrometer, MapQuant enabled us to search and find 18 more peptides based on m/z position and charge estimation. These 190 tryptic peptides cover 94.85% of the BSA sequence.

The data has shown evidence of linearity, at least for the highly abundant peptides observed in the range of 7 to 1600 fmoles. We have also developed a model for ionization efficiencies by calculating ionization coefficients for each amino acid. This model gives us the capability to describe the quantitation level of BSA as a whole protein. The applicability to quantitation of more complex mixtures such as a proteome appears to scale linearly with number of peptides, as long as the peak overlap density is kept at a low level.

With the capability of performing whole-cell proteome analysis (Lipton et al. 2002; Jaffe et al. 2004), a need to extend this process to quantitation has become increasingly apparent. Methods for measuring the state of an organism's proteome have been successful with the use of 2-D polyacrylamide gel protein maps, followed by spot excision, digestion of the protein with trypsin and peptide sequencing using reversed-phase liquid chromatography coupled to electrospray ionization mass spectrometry (ESI-MS) (Gygi et al. 1999; Pandey and Mann 2000). Quantitation of peptide mixtures using only chromatographic separation methods coupled to mass spectrometry has proven to be a more easily automated procedure than 2-D electrophoresis. However, quantitation of proteins in complex mixtures using the signal acquired from their constitutive tryptic peptides has become a very desirable and challenging goal. The reason being that the detection efficiency varies significantly from peptide to peptide; the molecular identities are not evident a priori and are dispersed unevenly throughout the multidimensional separations. Although, commercially available programs are available for quantitation, they are not designed for high-throughput proteomic data and the algorithms used in them are not publicly available.

Knowing the quantities of proteins in a biological system is crucial in understanding post-transcriptional events (Gygi et al. 1999; Futcher et al. 1999) including translational efficiency, post translational modifications, compartmentalization, interactions, and turnover.

In this study we proposed a data analysis method utilizing one or more chromatographic (or electrophoretic) separation dimensions and a mass separation dimension provided by the mass spectrometer. Data from an LC/MS experiment (hereafter referred to as *experiment*) can be analyzed after being formatted into a data-structure called a *2-D map*. A 2-D map is essentially a matrix whose rows and columns represent scans and m/z bins respectively. The 2-D map of a tryptic peptide from BSA is shown in Figure B1. The separation dimensions are considered orthogonal since they describe two independent properties of the peptides: mass and hydrophobicity. For this reason a parallelism can be drawn between a 2-D map and a 2-D electrophoresis gel, where in the latter the two dimensions represent mass and isoelectric point. The advantage of this visualization method is that the experimentalist gets a global view of the type of species eluting from the column and their relative position to each other.

Using the above as motivation we have developed open-source software, called MapQuant, which given large amounts of mass-spectrometry data, outputs quantitation for any organic species in the sample. We will discuss features of the software including several algorithms used in it. Furthermore, we have applied MapQuant in the study of BSA samples at different concentrations and tried to develop a peptide ionization model that would explain the abundances observed for the BSA tryptic peptides. Analyses of tryptic peptides of BSA have been carried out in the past (Bruce et al. 1999; Hirayama et al. 1990), however without any attempts for absolute quantitation using its constituent tryptic peptides. Another goal of this study was to set the ground on the standardization of storage and quantitation of mass spectrometry data and create a community where investigators can share their algorithms and data structures.

4.1.1. Data Acquisition

The sample used in this study was a BSA digest standard from Michrom BioResources (910/00002/15). Samples were injected, with a Famos^{TM} auto sampler, on a reversed-phase column coupled to a Finnigan LCQ^{TM} DECA XP+ mass spectrometer. The column was an in house made 15 cm x 75 μm capillary filled with Magic C_{18} resin.

The BSA digest was diluted in 95% water, 5% acetonitrile and 0.1% formic acid to a final volume of 10 μL for each of twenty-three experiments. Twenty-one of them involved seven different amounts of BSA peptides in triplicate. These BSA amounts include 7, 21.5, 66.67, 200, 500, 1600 and 5000 fmoles. The signal acquisition method for these experiments was carried out in the profile mode and did not involve any MS/MS scans, since they were meant for quantitation purposes (*q-experiments*). Moreover, these q-experiments were run from low to high concentrations to minimize carryover effects.

The remaining experiments involved two 5000 fmole samples that were run with a signal acquisition method that included MS/MS scans for sequencing purposes (*s-experiments*).

4.1.2. Data Storage

In order for MapQuant to be functional, a standard for storing raw LC/MS experiment data had to be developed. For the moment we are calling this standard OpenRaw but in the meantime we are in the process of developing an XML version of it. All OpenRaw files were created by a program written in Visual C++ using the API provided by the company Finnigan. OpenRaw file format has the advantage of being readable on all computer platforms due to its open-source nature.

The raw data of an experiment were stored in three functionally distinct folders which themselves were within a parent folder named after the LC/MS experiment (Figure B2). These folders are: (a) a global parameters folder, (b) an MS1 spectra archive folder and (c) an MS/MS spectra archive folder.

The global parameters folder holds four files. The *size.param* file stores information about the size of the data of an LC/MS experiment, i.e. total number of scans and total number of mass bins. The *RTSA.param* file, which stands for *Retention Time Sampl*ing *Array*, stores information about all the time points at which each mass spectrum was scanned. The *MSSA.param* file, which stands for Mass Sampling Array, stores information about the spacing of the sampling points in the m/z dimension given by the mass spectrometer. The file In*strumentMethod.param* stores information about the instrument method used by the user at that particular LC/MS experiment.

The MS1 spectra folder contains the file *expmnt_name.msar*. The extension *msar* stands for mass spectrum archive and, as the name indicates, it stores the ion-abundance signal from each mass spectrum in a concatenating manner.

Similarly, the MS/MS spectra folder stores the file *expmnt_name.ms2ar* which is a concatenation of the ion-abundance signal for all MS/MS spectra. MS/MS spectra can be analyzed for peptide identification by a sequencing program.

4.1.3. Data Analysis

Strategy: Identification – Quantitation

Since the mass accuracy of the spectrometer used in this study was not very high, we were reluctant to base peptide identification solely on the m/z of the peaks observed. Instead, we used the well established strategy of acquiring MS/MS spectra upon peptide

fragmentation, followed by sequencing using commercially available software, like SEQUEST (Eng et al. 1994). However, if we had chosen to perform MS/MS scans in our data acquisition scheme, we would have limited the number of MS scans that would have been acquired, hence reducing the sampling data points for quantitation. For example, a 140 min chromatographic run without acquiring any MS/MS spectra would yield 5881 scans, whereas the same run set to collect five MS/MS scans per MS scan would yield 4433 scans, of which about one fifth would be the MS scan used to reconstruct the 2-D map. To circumvent this problem we collected MS data with and without MS/MS spectra. These runs were referred to as *s-experiment*s and *q-experiments* respectively, as described in the data acquisition section. The strategy of how we linked the quantitative output from the q-experiment to the identification output of an s-experiment is outlined in Figure B3.

In order to link a sequenced MS/MS event to a fitted peak, the following algorithm was implemented, which was essentially encoded in the program ***assignq2.*** First, every entry in the SEQUEST summary file (Figure B3) reflects a sequencing event (MS/MS scan) which has a unique retention-time and m/z coordinates in a 2-D map, as indicated by the blue points in Figure B4. The sequence for these MS/MS scans that is depicted next to the blue points represent the highest scoring peptide given by SEQUEST. Isotopic clusters that were identified using MapQuant also have centroids, represented by red points in Figure B4. For each sequencing event a rectangular area that is 1 m/z and 80 scans wide in each dimension was searched for possible MapQuant peaks that lay within it. If there were multiple sequencing events that were assigned the same peptide, the peaks captured by these events were pooled and then ranked according to their Euclidean distance from their corresponding sequencing event. Also the charge predicted by MapQuant was checked for agreement with the charge used by the sequencing program. If the charge was different the peak was rejected. The procedure described above was used for all 21 q-experiments, thus creating a calibration table for further analyses.

Definitions and Data Structures

Experiment is the data structure that holds information about an LC/MS experiment. More specifically it holds information on the sampling of the eluent at different time points. Also it holds information about the sampling in the m/z dimension used by the spectrometer.

Scan is the sampling unit in the chromatography dimension

Mass bin is the sampling unit of the mass spectrometer when measuring the m/z of the produced ions.

Map2D is the data structure describing a 2-D map as defined in the introduction. It is stored in the form of matrix (Figure B5).

Mass spectrum is defined as the signal acquired at particular time point. It can be thought of as a row (or column) of the Map2D matrix (Figure B5).

Mass chromatogram is defined as the signal present at a fixed mass bin point. It can be also thought of as a row (or column) of the Map2D matrix (Figure B5).

Segment map is defined as a region of a parent 2-D map to which the operation of segmentation was performed. Segmentation is performed to partition the signal (peaks) in the map for easier analysis (Figure B6).

Peak is defined as a point in the 2-D map where it is considered to be a local maximum (as defined by a ***PeakF***in***der*** model)

FittedPeak (or ***FPeak***) is a peak that has been fitted to a particular model described by a mathematical equation, such as a two-dimensional Gaussian (Figure B6).

Peak group is the cluster of fitted peaks that can represent candidate co-eluting isotopic clusters (fig. 6).

Peak group map is defined as the minimum 2-D map needed for fitting the estimated number of isotopic clusters that a peak group might contain (Figure B6).

Isotopic cluster is a group of peaks that represent the isotopic variants of a molecular species (Figure 1).

MapQuant

MapQuant is a program designed for quantitation of data from an LC/MS experiment that is in OpenRaw format, as described above. The quantitation procedure involved formatting the data into a 2-D map and applying several algorithms on the 2-D map with the goal of outputting a list of abundances of possible isotopic clusters. Below, we are presenting the order that several algorithms were applied to the data acquired from the BSA samples mentioned in the data acquisition section.

Since the data were quite noisy, especially in the chromatography dimension, noise filters were applied. More specifically, smoothing algorithms such as applying a moving-averaging filter were applied (Press 1992). The purpose of the smoothing algorithm is to facilitate the detection of all local maxima (peaks) found in the 2-D map.

The second step in the data processing involved segmentation of the 2-D map into smaller areas called segments, such that overlapping peaks were confined into unique segment maps (Figure B7). The purpose of the segmentation algorithm was to minimize the number of data points being used simultaneously for fitting. It also made sure that overlapping peaks were included in the same fitting iteration.

The third step of the quantitation process involved the application of a peak detection algorithm for finding the positions of the local maxima in every segment map, and then using that information as initial conditions for the curve fitting algorithm.
The fourth step in the procedure involved single linkage clustering of the fitted peaks into clusters referred to as peak groups. Peak groups represent co-eluting peaks that might include one or more possible isotopic clusters.

Finally, more refined fitting was performed for each peak group. This refining algorithm involved an iteration of subtraction and residual fitting based on previously estimated peak widths. This algorithm was chosen because peptides that had charges of +2 and +3, had isotopic peaks that were significantly overlapping. However, peptides with a +4

charge were impossible to resolve by this method, since the resolution offered by the Finnigan LCQ^{TM} DECA XP+ spectrometer was too low.

The above five steps were included in the structure of scripts called *MQScripts* that can be read by MapQuant's parser called *MQParser*. MapQuant algorithms were written in ANSI C and the MQParser was written using the general-purpose parser generator program called *bison*. Moreover, MapQuant includes minimal visualization features including 2-D map visualization as well as mass spectrum and mass chromatogram visualization.

In the next section, some of the most important functions recognized by the MQParser are discussed in detail, along with the algorithms they encapsulate.

Algorithms

1. Smoothing by convolution
2. Watershed segmentation
3. Peak Finding and Peak Fitting
4. Peak Clustering
5. Refine m/z peaks (Deconvolving by fitting and subtracting)
6. Deconvolve and isotopic carbon peaks

Smoothing by convolution

Convolution is the equivalent of band filtering in the frequency domain after a Fourier transform has been applied to either a mass chromatogram or a mass spectrum. The implementation of this filter utilizes the mathematical property shown in equation 1 (Press 1992), where s is the signal array and h is the impulse array. S and H are the Fourier transforms of s and h respectively.

$$s \otimes h = F^{-1}(S.H) \qquad \textbf{(1)}$$

Related Functions in MQScript

Map2D **Map_ApplyFilter** (Map2D *map*, char* *filter*, double *dim*)

The string *filter* describes the impulse function that will be applied to each mass spectrum or mass chromatogram of the 2-D map "*map*". The value of *filter* can either be "BC_*n*" or "GS_*n*_*m*" where *n* and *m* are integers. BC refers to a box-car filter (otherwise called moving-average filter) of *n* points and GS refers to a Gaussian filter of standard deviation *n.m*. The value of *dim* can be either RT_DIMENSION or MZ_DIMENSION, referring to the dimension of the 2-D map that the filter will be applied.

Watershed segmentation

The goal of this quantitation analysis is to fit every peak in the 2-D map. However, since fitting all peaks at the same time is computationally too expensive, a strategy involving compartmentalization is considered. This is achieved by segmenting the map using the watershed segmentation algorithm (Vincent and Soille 1991).

Related Functions in MQScript

Map2D **Map_Segment**(Map2D *map*, double *b*in_*size*)

This function takes the 2-D map *map* and segments it using the watershed algorithm. The parameter *b*in_*size* is used to denote the stepping size used by the flooding procedure of the watershed algorithm.

VariantArray **Map_GetSegmentArray**(Map2D *map*)

This function returns an array of type 'segment' which defines the boundaries of a segment map.

Map2D **Segment_GetMap**(Segment *segment*, Map2D *source*, Map2D *mask*)

This function extracts a segment map from the map *source*, using a map which acts as a mask. Each point in the 2-D map *mask* holds the segment number in which it belongs to. This procedure is illustrated in Figure B7.

Peak Finding and Peak Fitting

After compartmentalizing the parent 2-D map into segment maps, the goal of peak finding and peak fitting becomes computationally more manageable. The peak detection algorithm used here uses concepts from mathematical morphology such as a structuring element (Ritter and Wilson 2001). A structuring element can be considered a small binary image that an image operator can take as input along with the image of interest. An example would be the image operation of erosion (symbolized by ^) between the image-signal *S* and the structuring element *N* that would act as a kernel (analogous to the response element used in convolution). The operation mentioned above is shown mathematically in equation 2.

T = S ^ N **(2)**

The peak detection algorithm uses a structuring element such as the ones in Figure B8, in order to decide which neighboring points for each data point in the 2-D map are to be included in the image operation. The shaded point indicates the point of reference of the structuring element.

In the operation of peak-detection, a data point in the 2-D map is considered a local maximum only if its value is larger than all the neighboring points defined in the structuring element as 1. Mathematically speaking

$$a_k = \begin{cases} 1 \text{ if } \{\sum_i \Lambda(s_k, N_i)\} = |N| \\ 0 \text{ otherwise} \end{cases} \qquad \textbf{(3)}$$

Where $\Lambda(p,q) = \begin{cases} 1 \text{ if } p >= q \\ 0 \text{ otherwise} \end{cases}$ and $|N| = \sum_i N_i$

N represents the structuring element and s_k the value of the data point k in the 2-D map S.

To avoid detecting pseudo-peaks due to the noisy signal an abundance threshold is also set for all the points in the structuring element. The abundance threshold is usually set to m + 2s, where m is the median of the map and s is the average absolute deviation from the median.

After a list of candidate peaks is formed, they are fitted as a sum of curves described by a mathematical equation. For example, if in a segment map there are n candidate peaks, and if each peak is chosen to be fitted as curve C, then the whole segment-map would be fitted as $\sum_{i}^{n} C_i$. In this study we chose to fit each curve with a double Gaussian, referred to from now on as the gaussioid curve, i.e.

$$f(m,r;A,r_{o,}m_o,\sigma_m,\sigma_r) = \frac{A}{2\pi\sigma_m\sigma_r} e^{-\frac{(r-r_o)^2}{2\sigma_r{}^2}} e^{-\frac{(m-m_o)^2}{2\sigma_m{}^2}} . \quad \textbf{(4)}$$

As seen from the equation, the number of parameters to be fitted per peak is five, i.e. abundance A, retention-time centroid r_o, mass-over-charge centroid m_o, the standard deviation of the Gaussian in the retention time dimension σ_r and finally the standard deviation of the Gaussian in the mass-over-charge σ_m. The method used for peak-fitting is the "non-linear least squares" method (Press 1992). It is a minimization method using steepest descent. It requires knowledge of the first derivative for each of the parameters to be fitted. For example if there are n candidate peaks in a segment map and we want to fit them using the gaussioid curve, we would need to fit 5n parameters and the algorithm would require 5n partial derivatives.

Related Functions in MQScript

Segment **Segment_FindAndFitPeaks**(Segment *segment*, Map2D *map*, char* *structel*, double *abuthr*, double *noisefactor*, double *numberofsd*, char* *curve*)

This function finds and fits peaks in the 2-D map *map*, and returns them inside a variable of type 'Segment'. The arguments used by the peak-finding part of the algorithm are *structel* and *abuthr*. The argument *structel* refers to the structuring element used in the peak-finding of the algorithm. It can take values such as "N9x3E", "N3x3R", etc (Figure B8). The argument *abuthr* denotes the threshold value which needs to be met by all the points in the structuring element. The argument *curve* refers to the peak-fitting part of the algorithm and it denotes the type of curve that the peak should be fitted too. It can take values such as "NR_GAUSSIOD" or "NR_EM_GAUSSIOD". The arguments *noisefactor* and *numberofsd* are not used anymore.

double **Map_GetMedian**(Map2D *map*) and

double **Map_GetAvgAbsDevFromMedian**(Map2D *map*)

The function Map_GetMedian calculates a distribution of the intensities of all the points in the 2-D map *map* and returns its median. Similarly Map_GetAvgAbsDevFromMedian returns the average absolute deviation from the median.

Peak Clustering

In order to decide which fitted peaks belong to which isotopic clusters we cluster the peaks into data structures called peak groups. Peaks belonging in an isotopic cluster have restrictions in their relative position in the isotopic cluster. More specifically isotopic peaks have to be co-eluting (Figure B1). Secondly, the peaks cannot be separated more than 1 m/z unit, which is maximum distance defined by peptides with charge +1. The above natural restrictions, co-elution and the 1 m/z maximum distance limit, are taken into account by limiting the number of inter-peak distances when peaks are clustered.

Related Functions in MQScript

double **FPeakFile_ClusterAndSave**(char* *sz*In*Filename*, char* *szOutFilename*, double *scan_thr*, double *mz_thr*, double *abund_thr*, double *nL*in*kageType*)

This function reads the fpeaks from the file *sz*In*Filename*, clusters them into peak groups and then outputs them into the file *szOutFilename*. The arguments *scan_thr* is used to set the scan tolerance of the co-elution restriction mentioned above. Likewise the argument *mz_thr* is used to set the m/z tolerance for the maximum allowed distance between two isotopic peaks. This value was set to 1.1 m/z, and calculated from known +1 peptides.

Refine m/z peaks (Deconvolving by fitting and subtracting)

Due to the fact that the spectrometer cannot resolve isotopic peaks of peptides that have charges of +3 and above, we apply an algorithm following a 'fit and subtract' strategy. We use a simple test for discriminating between 'oversized' or 'undersized' peaks from 'normal' peaks. The metric applied in this case makes the assumption that peak width in the m/z dimension is constant throughout the experiment. The pseudo code for this algorithm is provided in the supplementary material.

Related Functions in MQScript

PeakGroup **PeakGroup_Refine6**(Peakgroup *peakgroup*, Map2D *PGMap*, VariantArray *pPGMapFPeaks*, char* *structel*, double *numberofsd*, char* *curve*)

Deconvolve and fit isotopic carbon peaks

In this algorithm we assume that each peak group may represent one or more isotopic clusters. So we devise an algorithm for sequential prediction of each isotopic cluster by going through the fpeaks based on increasing m/z. The peaks that are candidates for an isotopic cluster are to be fitted using a binomially distributed sum of two-dimensional Gaussians.

$$f(m,r;A,r_{o,}m_o,\sigma_m,\sigma_r,c,z) = A\sum_i \frac{B(i;c,p)}{2\pi\sigma_m\sigma_r} e^{-\frac{(r-r_o)^2}{2\sigma_r^2}} e^{-\frac{(m-(m_o+\frac{i}{z}))^2}{2\sigma_m^2}} \quad \textbf{(5)}$$

Where $B(i;c,p) = \binom{c}{i} p^{c-i}(1-p)^i$ **(6)**

Related Functions in MQScript

PeakGroup **PeakGroup_Deconv3**(PeakGroup peakgroup, Map2D PGMap, VariantArray pPGMapFPeaks, char* curve);
The above function can be divided into two parts which are iterated until all fpeaks are distributed into isotopic clusters:

1. Guess the most likely subset of the fpeaks in a peak group, which can comprise a possible isotopic cluster.

Substitute those peaks with a binomially distributed gaussioid curve (denoted with the string "NR_BD_GAUSSIOD") and fit the peak group map with equation 7, where m is the number of single gaussioid curves C (eq. 4) and n is the number of binomially distributed gaussioid curves B (eq. 5).

$$\sum_{i}^{m} C_i + \sum_{i}^{n} B_i \qquad \textbf{(7)}$$

First we provide a description of the first part of the algorithm which itself is comprised of two steps. As a first step it involves the finding of all possible sets of fpeaks whose m/z forms an arithmetic series, as shown in equation 8, and hence might provide evidence of belonging to the isotopic variant peaks of a charged molecular species. This algorithm is referred to a *charge deconvolution algorithm.*

$$\begin{aligned} & a_o \\ & a_1 = a_o + d \\ & a_2 = a_o + 2d \\ & a_3 = a_o + 3d \\ & \ldots \end{aligned}$$

where d = 1/Z , Z is the charge of the molecular species **(8)**

Figure B9 shows a possible arrangement of fpeaks in a peak group and the schematic description of the charge deconvolution algorithm. This algorithm extracts sets of fpeaks that could belong to candidate isotopic clusters. As seen in Figure B9, for each fpeak belonging to the peak group, which itself can be represented as a directed acyclic graph, a search is being performed through it and stored in a tree. For example, steps 2-4 describe the creation of a tree whose root is peak A, step 5 describes the creation of the tree whose root is peak B, etc. Each tree stores all possible paths that link the root with every leaf. The paths in all the trees represent all candidate isotopic clusters for the peak group under study. In order to choose the most plausible isotopic cluster candidate we rank these isotopic clusters according to the sum of the abundances of their constituent peaks. In this way, low abundant molecular species do not interfere with the deconvolution process since they are deconvolved later. The next step after charge is to take the most abundant candidate isotopic cluster discussed above, and to try to estimate a possible carbon content by comparing the abundances of its comprising peaks to the binomial distribution described in equation 6. The metric used in this selection is the average square deviation form an idealized, hence calculated binomial distribution for a particular carbon content.

Similar algorithms for carbon deconvolution have been reported in the literature (Wehofsky et al. 2001).

Finally, after the first part of the algorithm is called iteratively and the number of isotopic clusters represented in the peak group is found, they are fitted with equation 7.

4.2. Results and Discussion

Coverage

The BSA sequence used in the study was identified by running SEQUEST on nine BSA sequences extracted from the National Center for Biotechnology Information (NCBI) database. From the peptides that scored (xcorr > 2.0) it was evident that the 24 amino acid leading peptide was not present in the mature form of the BSA used in the experiment, implying a protein of 583 amino acids in length. The sequence is shown in Figure B10 and referred to from now on as mBSA-A214T.

SEQUEST was re-run using a protein database that included mBSA-A214T as well as 58 trypsin sequences. Enzyme settings were set to none, in order to increase the confidence for the identity of the tryptic ones. The program was set to search for +1, +2 and +3 charged variants. Moreover, it was set to take into account possible amino acid modifications such as lysine and arginine carbamylation, methionine, histidine and tryptophan oxidation, and glutamine N-terminal loss of ammonia. The following charge specific *xcorr* cutoff values were set: 3 for +3 peptides, 2 for +2 peptides and 1 for +1 peptides. The peptide charge variants that passed the charge-specific cross-correlation score thresholds amounted to 242. From these peptide charge variants only half (121) were fully tryptic on both termini. By observing sequences of the non-tryptic peptides we concluded that an enzyme with chymotrypsin activity was present in the digestion mixture, since 74 of the amino acids found at the C-terminus of the restriction site were either phenylalanine, tyrosine or leucine (Antal et al 2001). However we hypothesize that chymotrypsin activity was attributed to an enzyme that was co-purified with trypsin and in a lesser amount, since only 8 peptide charge variants were fully chymotryptic.

Out of the 242 BSA peptide charge variants 70 were regarded as false positives and were rejected because no signal was found on the two-dimensional maps. The remaining 172 peptide charge variants cover 540 amino-acid residues out of a total of 583 which corresponds to 92.62%.

MapQuant Performance In order to evaluate MapQuant's performance we try to estimate the percentage of SEQUEST hits that have been linked to a possible MapQuant isotopic cluster of peaks. It should be noted that the program performs better with +2 peptides. An obvious reason for the bias mentioned above is due to the low resolution of the spectrometer. For example the average m/z bin size was about 1/15 (0.067) m/z units wide. This means that peaks belonging to an isotopic cluster of a +3 peptide would be only 5 bins apart given an average peak width of 0.16 m/z (2.4 bins). This makes it extremely difficult for any algorithm to resolve the peaks.

Moreover, it should be noted that the lack of finding an isotopic cluster with the correct charge does not imply that the algorithm did not find any peaks that were in the vicinity of the MS/MS event. Another reason for the lack of finding all peaks can be attributed to the way the two experiments (q-experiment and s-experiment) were chosen to be aligned and to the size of the window chosen to be searched around the MS/MS events. Thirdly, since only certain peptides give high enough signal to be detected in the high concentrations of the sample, we assume that a significant number of the total number of peptides would be detectable in the low concentration data points. For this reason we observed that the percentage cover increased for the four most concentrated samples (12 points).

Additional Non-SEQUEST Peptides In addition to the 172 peptide charge variants we included 18 peptide charge variants that were not found by SEQUEST. There are two possible reasons for peptide isotopic clusters present on a 2-D map not to be identified by SEQUEST. One reason could be the fact that an MS/MS spectrum corresponding to a peptide isotopic cluster is not interpretable by the program. Another reason could be the complete absence of MS/MS spectra for a peptide isotopic cluster due to the difficulty of sampling MS/MS spectra for a 2-D map densely populated by peaks.

These 18 non-SEQUEST peptides were divided into four groups based on the criteria for accepting the identity to be true. The first group included seven peptides that were found by MapQuant after analyzing a 2-D map that had been collected on a high-resolution Fourier Transform Ion Cyclotron Resonance Mass Spectrometer). A second group included 7 peptides that were found manually due to the presence of other same-sequence charge variants that co-eluted with them. The third group included three small non-overlapping tryptic peptides of charge +1, that were covering parts of the protein that were not covered by all previous peptides. Finally, the last group included a peptide which had to be fitted manually because of its marginal position in the 2-D map, as it had the lowest m/z of all peptides.

These new non-SEQUEST peptide charge variants increased their total number to 190 and also the sequence cover of BSA to 94.85%.

Amino Acid Modifications and N-Terminal Cyclizations
In this study we were also interested in finding out possible modifications. Among the 172 peptides examined, we focused on the following modifications, mainly on S-carboxymethylation of cysteines, on the oxidation of methionine and histidine, on carbamylation of lysine and arginine and on neutral loss of ammonia from N-terminal glutamine. These four modifications were the ones set for SEQUEST to search for. In addition, the 19 non-SEQUEST peptide charged variants described above, include two cases where we observe neutral loss of ammonia and one case of arginine modification.

Out of the eight cases representing neutral loss of ammonia, all of them correspond to peptides whose sequences have the amino acid glutamine at their N-terminus. This is consistent with the formation of pyroglutamate reported in the literature (Baldwin et al. 1990).

With regard to carbamylation, 5 peptides were observed to have carbamylated lysines, and 4 peptides were observed to have carbamylated arginines (supplementary table 3). Peptide K*QTALVELLK indicated that lysine-548 was carbamylated. Moreover, lysine-548 is also known to be glycated (Wada 1996), indicating a sequence hot spot for attack by acidic molecules.

Linear response

We also wanted to investigate the range of linear response of an ESI mass spectrometer, such as the Finnigan LCQ^{TM} DECA XP+. Although, the results pertain to the particular instrument used in this study, our long-term goal is to be able to use the BSA tryptic peptide mix as a calibration standard that has to be run on the instrument where the proteomic study will occur.

In this study we focused on the five most abundant peptides, which had been confirmed by SEQUEST and had a charge of +2. The theoretical number of data points for these 5 peptides is equal to 105. For the graphs shown in Figure B11, the abundances of 72 isotopic clusters were used, which included 59 automatically found by MapQuant out of which 8 had to replaced by manually fitted isotopic clusters, and 13 completely new manually fitted isotopic clusters.

As we see from Figure B12, when we plot the logarithms of fmoles vs. ion volume units for each peptide charged variant we see that the 5000 fmole point is an outlier as it shows evidence of saturation. For this reason we use the data points from 7 fmoles to 1600 fmoles for fitting. The curve used to fit the data points mentioned above is of the form $y = Ax^n$. The correlations observed for this log-log regression are very close to 1. Moreover, parameter n, which represents a measure of deviation from linearity, is very close to one for all peptides, especially for R.RPC#FSALTPDETYVPK.A (1.05439) and R.KVPQVSTPTLVEVSR.S (1.07880). However, further investigation for the reasons of the variation of parameter n should be considered in the future.

Model for Ionization

Finally, we thought that one aspect of mass-spectrometry that MapQuant could be used to explore, is modeling the variability of ionization of different peptides. The approach we took was to cluster the 190 peptide charged variants into 39 non-overlapping peptide clusters, shown by a double line in Figure B10. The abundance for each peptide cluster was calculated as the sum of the abundances for all its constituent peptide charged variants. This was done only for the 1600 fmole data point since it is the most abundant data point in the linear region. In this way we can assume that the number of molecules introduced into the mass spectrometer per peptide cluster is equal and thus allowing us to proceed into formulating a model for explaining why the signal acquired per peptide cluster is not equal to each other.

In this study we developed a linear model to explain the ionization variation of each peptide cluster, where each amino acid contributes either negatively or positively to ionization efficiency. Similar models have been applied to the prediction of the retention time of peptides in reverse phase columns (Chabanet and Yvon, 1992). Since it is a linear model, the fact that we sum the abundances of the constituent peptide charged variants for each peptide cluster is compatible with the model.

In order to calculate the coefficients of ionization for each amino acid we use the equation 9, where **X** is the matrix amino acid composition for the 39 peptide

$$Y = X\beta + \beta_o \qquad \textbf{(9)}$$

clusters and β is 1 x 20 vector holding the ionization coefficient for each the twenty known amino acids. In the case of **Y**, we assume that there exists a maximum ionization value for the peptide clusters that ionize well. For this reason we normalize the abundance of each peptide cluster by the maximum and use that as the **Y** vector in the regression. The regression is performed using the statistical toolkit in MATLAB and in particular with the function *regress*. The correlation of this regression is 0.7967, with as p-value of 4.8×10^{-3}. The regression was also cross validated using the leave-one-out method with the correlation value of 0.5476.

Using the ionization coefficients, we can see that for a significant number of amino acids, their ionization coefficients are positive. Ionization coefficients for methionine and tryptophan can be considered the least reliable since the BSA sequence contains only 4 and 2 respectively. Only six amino acids have ionization coefficients that are less than 0.01. At least for two of them, i.e. proline and phenylalanine hypotheses could be formulated as to explain their negative ionization coefficients. Although, the exact mechanism of ionization in the gaseous phase is not known (Cole 2000), the ionization coefficients of these two amino acids point to an ionization model which reflects amino acid ionization properties that are known for the aqueous phase.

More specifically, proline which has the second most negative ionization coefficient is the only proteinogenic amino acid that forms a tertiary amide when it is part of a peptide. This means that it does not have a hydrogen on the amide group and can therefore not act as a hydrogen bond donor. Another amino acid that has a negative ionization coefficient

is phenylalanine. Phenylalanine is the only proteinogenic amino acid whose side chain is both aromatic and completely non-polar. It is also known that aromatic amino acids contribute negatively to the pK_a of acidic and basic groups, which reflects ability to be protonated.

Moreover, no correlation was found between amino acid ionization coefficients and either isoelectric points or hydrophobicity coefficients. As far as modifications and N-terminal cyclizations are concerned we did not distinguish, for example, between glutamine and pyroglutamate and neither between carbamylated and uncarbamylated lysines or arginines. It is evident, that these regression results call for the collection of more datasets similar to this one, but for different proteins, in order to improve the correlation of the regression and the confidence of the ionization coefficients.

Concluding Remarks

One purpose of this study was to promote community sharing and standardization of quantitative mass spectrometry. We believe that MapQuant is an excellent beginning for the above goal as it is an open-source and versatile software. An XML version of the OpenRaw data format will facilitate the exchange of raw mass spectrometry data. Furthermore, we have ensured that MapQuant can be compiled and run on both Windows and Linux platforms. Another advantage of MapQuant is that it can process data acquired on a Fourier Transform Ion Cyclotron Resonance Mass Spectrometer (FTICR-MS).

In conclusion, our goal was to be able to apply the techniques used for BSA to quantitate proteins in complex mixtures such as proteomes of microorganisms that have small genomes. In this study we show that tryptic peptides from BSA can behave linearly and ionize according to a linear ionization coefficient model postulated above. We believe that BSA or other proteins for that matter can be used as standards either external or internal for calibration of different types of mass spectrometers, including ones that use quadrupole ion traps (QIT), linear traps or cyclotron resonance cells.

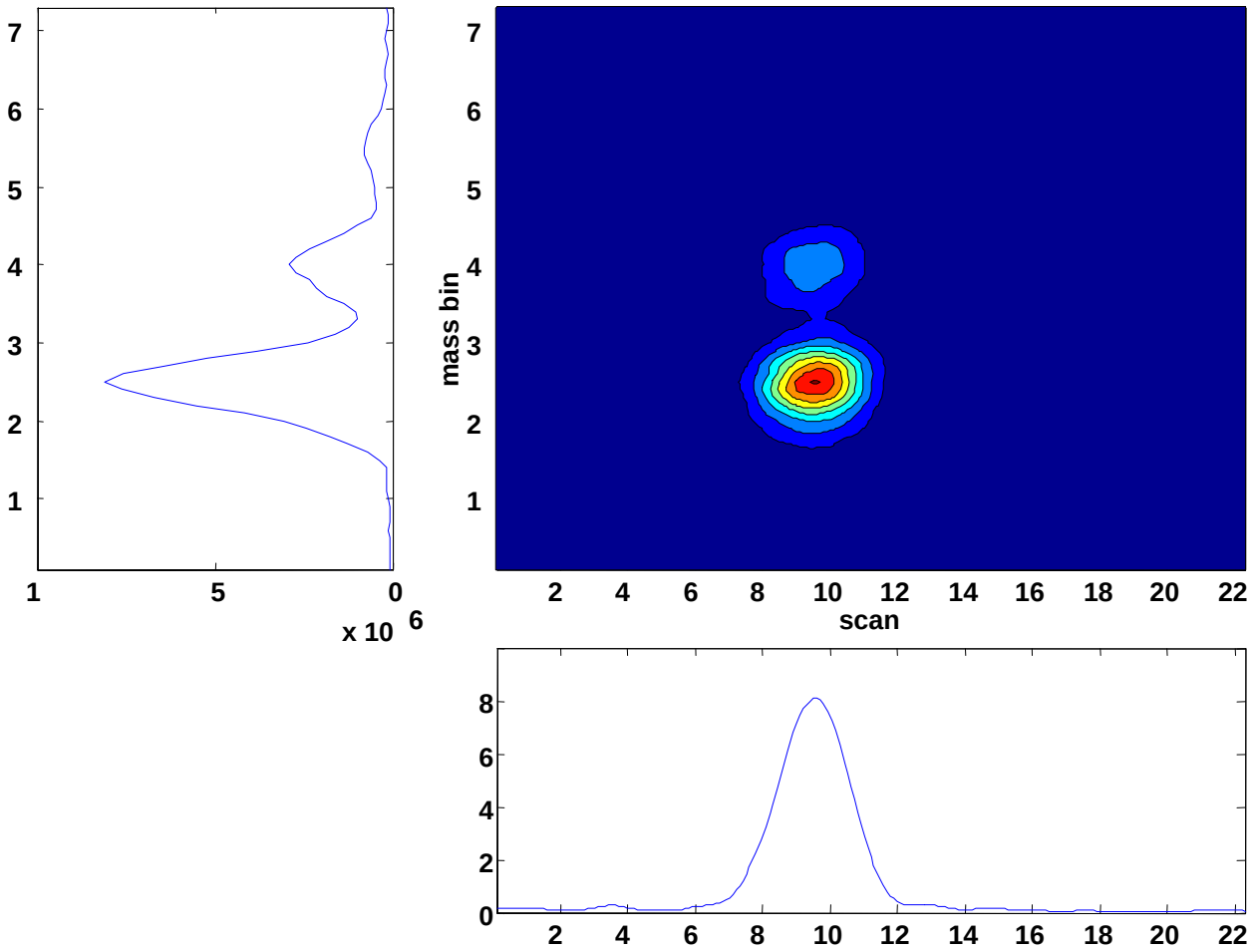


Figure B1. A detailed look at an isotopic cluster as it is visualized in a 2-D map.

Sections of the 2-D map, such as relevant mass spectrum and mass chromatogram are also shown.

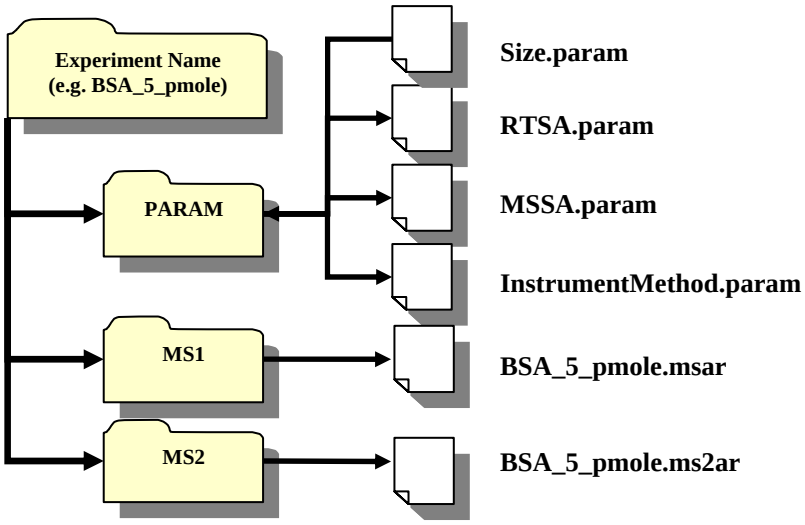


Figure B2. The format (file tree structure) used by MapQuant to store raw data

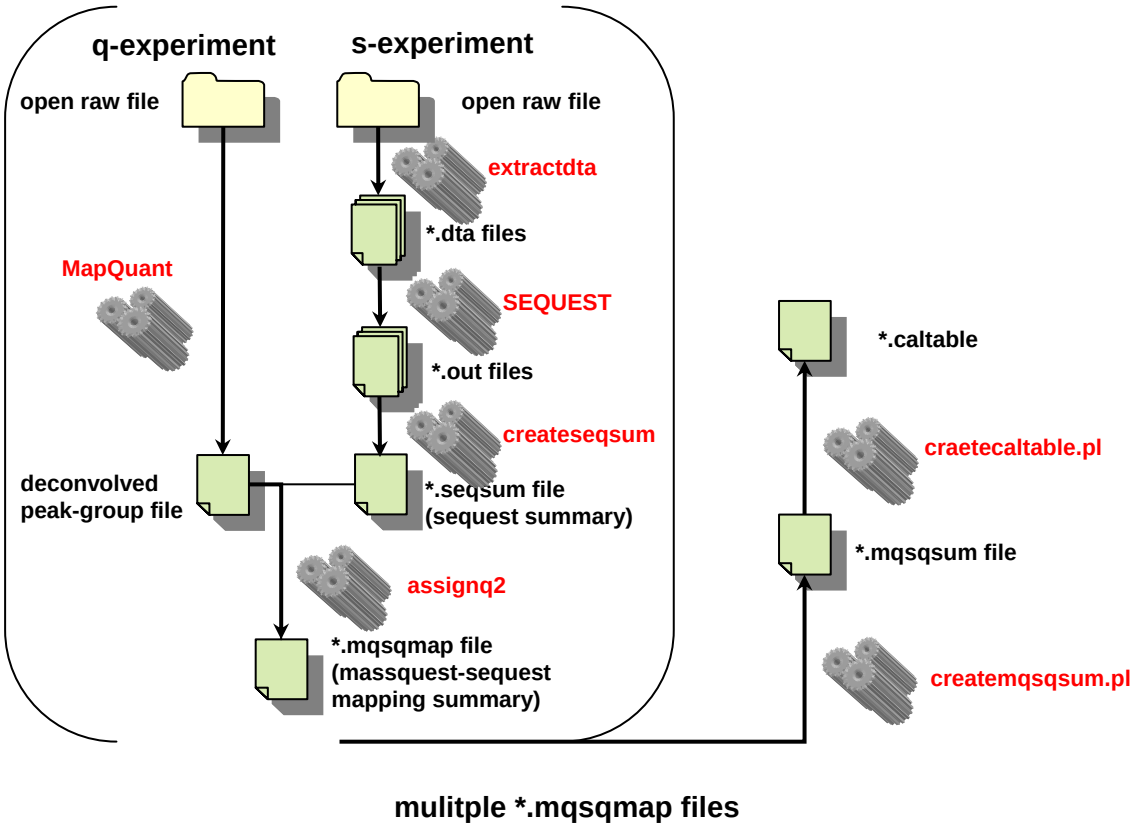


Figure B3. The pipeline employed to link an s-experiment with a q-experiment.

Regarding the q-experiment the program presented in this report, MapQuant, is used to extract fitted peaks from the raw data. On the other hand, SEQUEST with the help of the program *extractdta*, provides identification information for each dta file. Then all

SEQUEST output files are summarized in one summary file using the script *createseqsum.pl*.

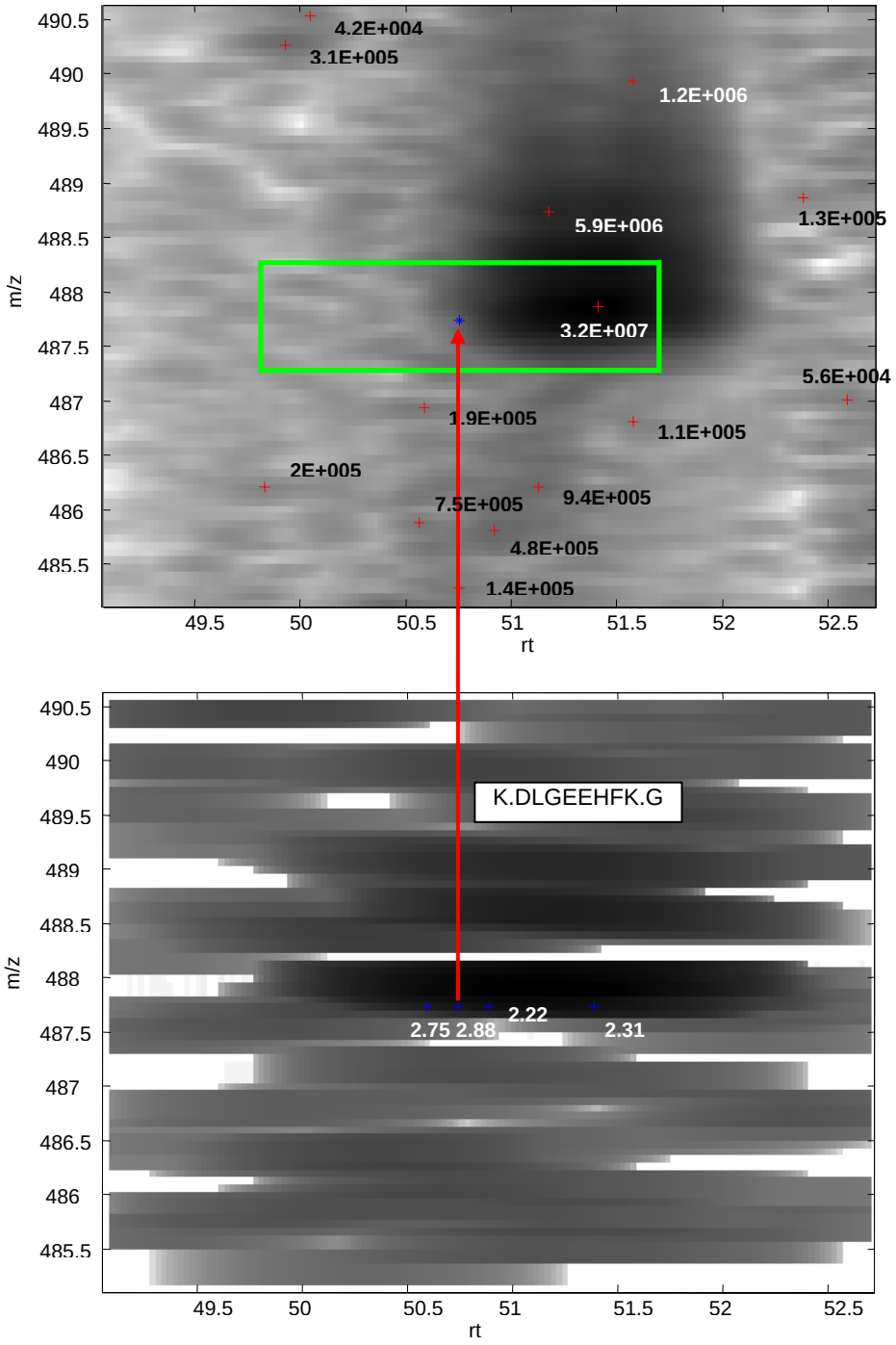


Figure B4. Mapping between sequenced MS2 events and quantitated isotopic clusters.

MS2 events, marked by blue crosses, correspond to the peptide sequence K.DLGEEHFK.G, as shown in the 2-D map extracted from an s-experiment (lower map). For each MS2 event a 160-scan x 1-m/z search window (in this illustration, 40 scans on each side in the RT dimension and 0.5 m/z on each side in the m/z dimension) is drawn in the 2-D map extracted from a q-experiment (upper map). Quantitated isotopic clusters using MapQuant are marked by red crosses and are labeled by their integrated intensities. Any isotopic clusters found within the windows defined by the MS2 events are assigned to them.

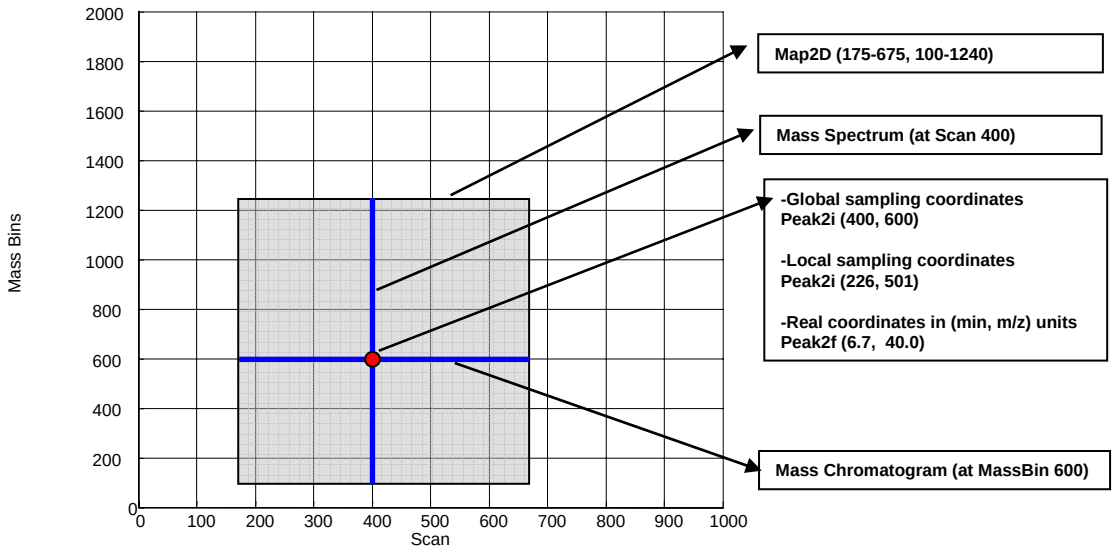


Figure B5. Illustration of the definitions surrounding the concept of a 2-D map.

As seen in the figure, a 2-D map can also describe a fraction of an experiment, indicated by the shaded rectangle. A 2-D map is defined by scan boundaries (e.g. 175 – 675) and by mass bin boundaries (e.g. 100 – 1240). Any column of a 2-D map is defined as a *mass spectrum* at a particular scan, and any row is defined as a *mass chromatogram* at a particular mass bin. Positions of data points in a 2-D map can be addressed in three different ways: (a) Using global sampling coordinates, where position is given in scan and mass bin units that refer to the experiment as a whole, (b) using local sampling coordinates, where position is given in scan and mass bin units but using a local frame of reference and (c) using real number coordinates, where position in the 2-D map is described in units of the physical quantities that the sampling points refer to, i.e. minutes and m/z.

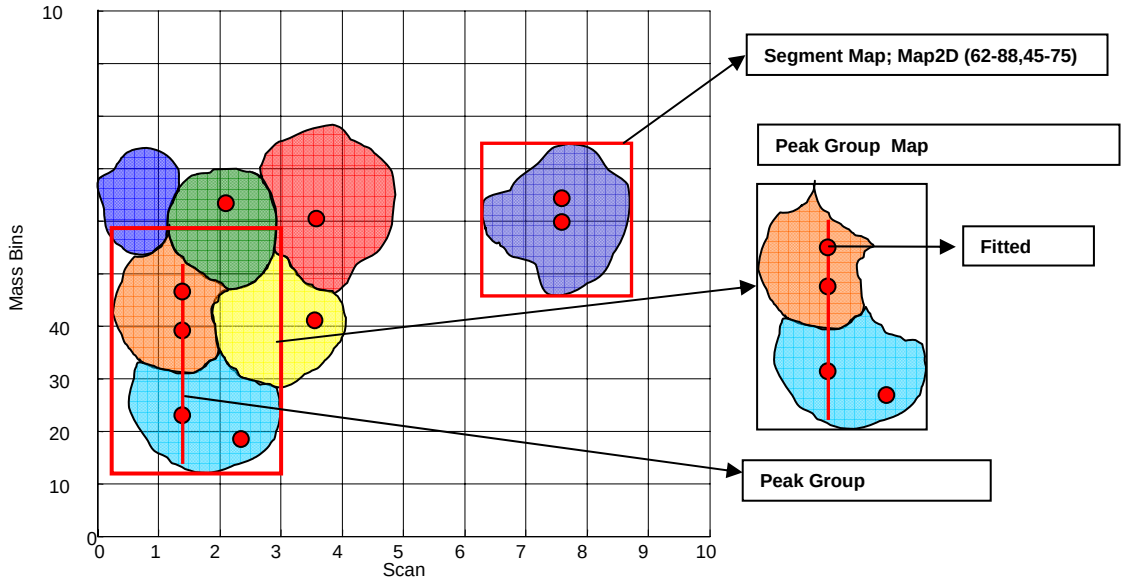


Figure B6. Illustration of data structures and concepts required for the understanding of the algorithms.

A *segment map* is a 2-D map that contains all the data points belonging to a segment as result of performing the operation of watershed segmentation on a parent 2-D map. A *peak group* is defined as a cluster of fitted peaks that can represent candidate co eluting isotopic clusters. A *peak group map* is the minimum 2-D map needed for fitting the estimated number of isotopic clusters that a peak group might contain.

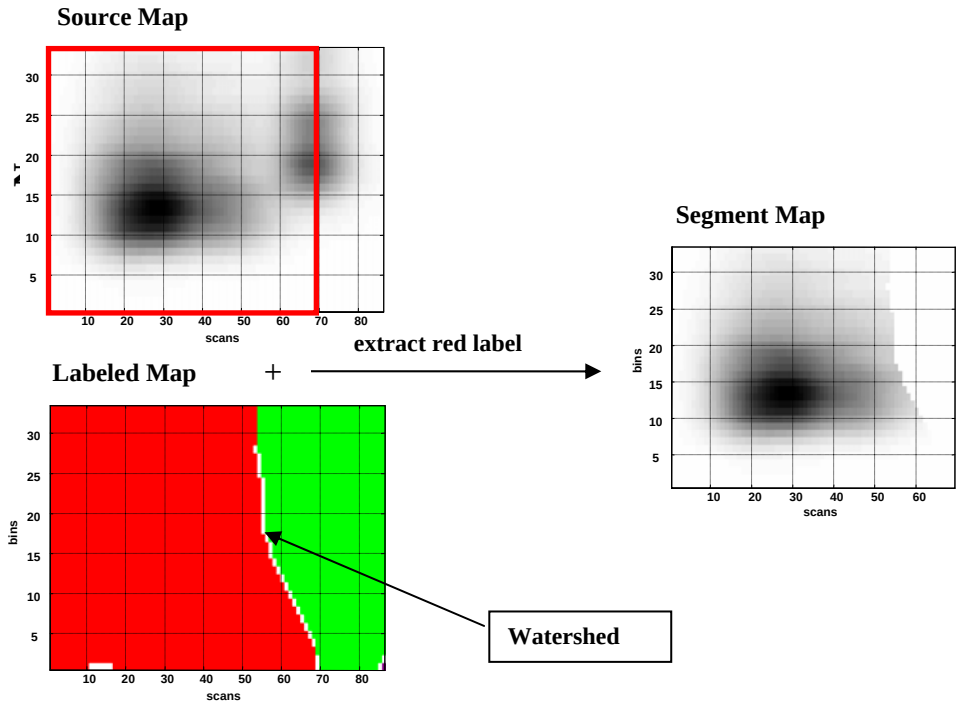


Figure B7. Operation of watershed segmentation on a 2-D map.

This algorithm is utilized to divide the map in non-overlapping regions so that fitting individual peaks becomes less computationally intensive. The product of segmentation is a 2-D map called *labeled map* where each data point is given a segment number which it belongs to (indicated by different shades). The labeled map can be used as a guiding mask to extract the data points needed for a particular segment, thus creating a segment map as described in Figure B5.

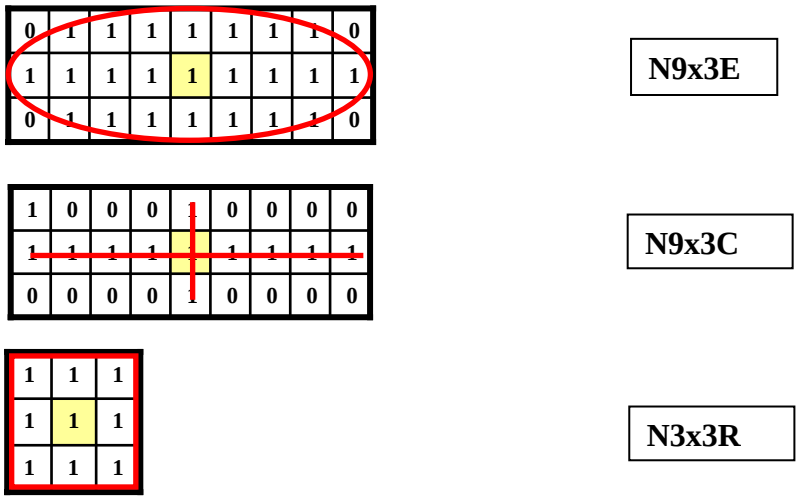


Figure B8. A collection of different structure elements used for different map operations, such as opening, closing and peak finding.

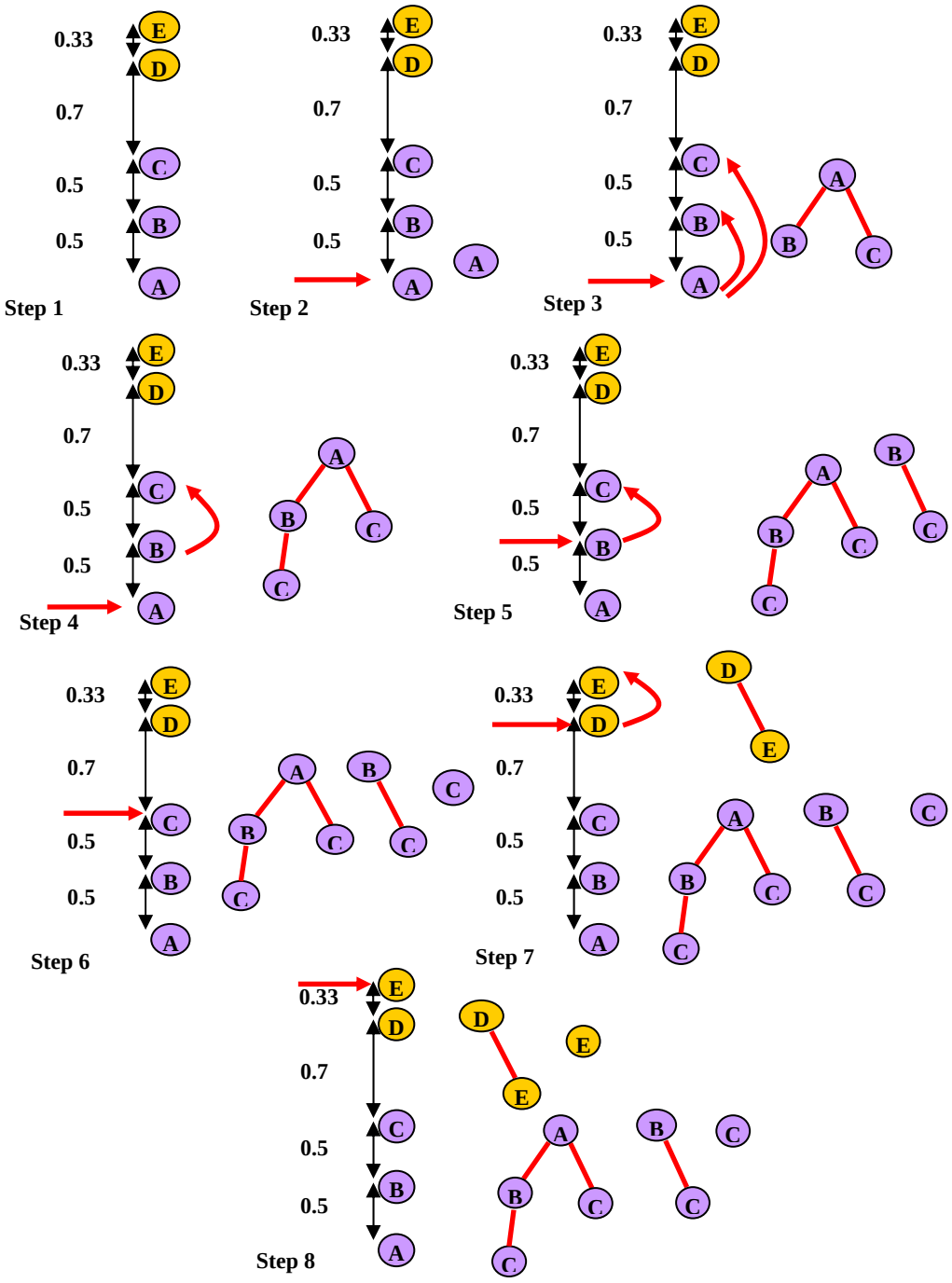


Figure B9. Schematic way description the charge deconvolution algorithm.

The algorithm's goal is to go through all peaks that belong to the peak group and construct a tree for each one of them, which will represent all possible equidistant groups of peaks that might be candidates for defining an isotopic cluster. For building the first tree, the algorithm starts from the peak with the lowest m/z (peak A, step 2) and searches and picks all the peaks (peak B and C) that in relation with the root peak (peak A) have

an m/z distance that is compatible with a possible charge state of an isotopic cluster (+2 for peak B and +1 for peak C.

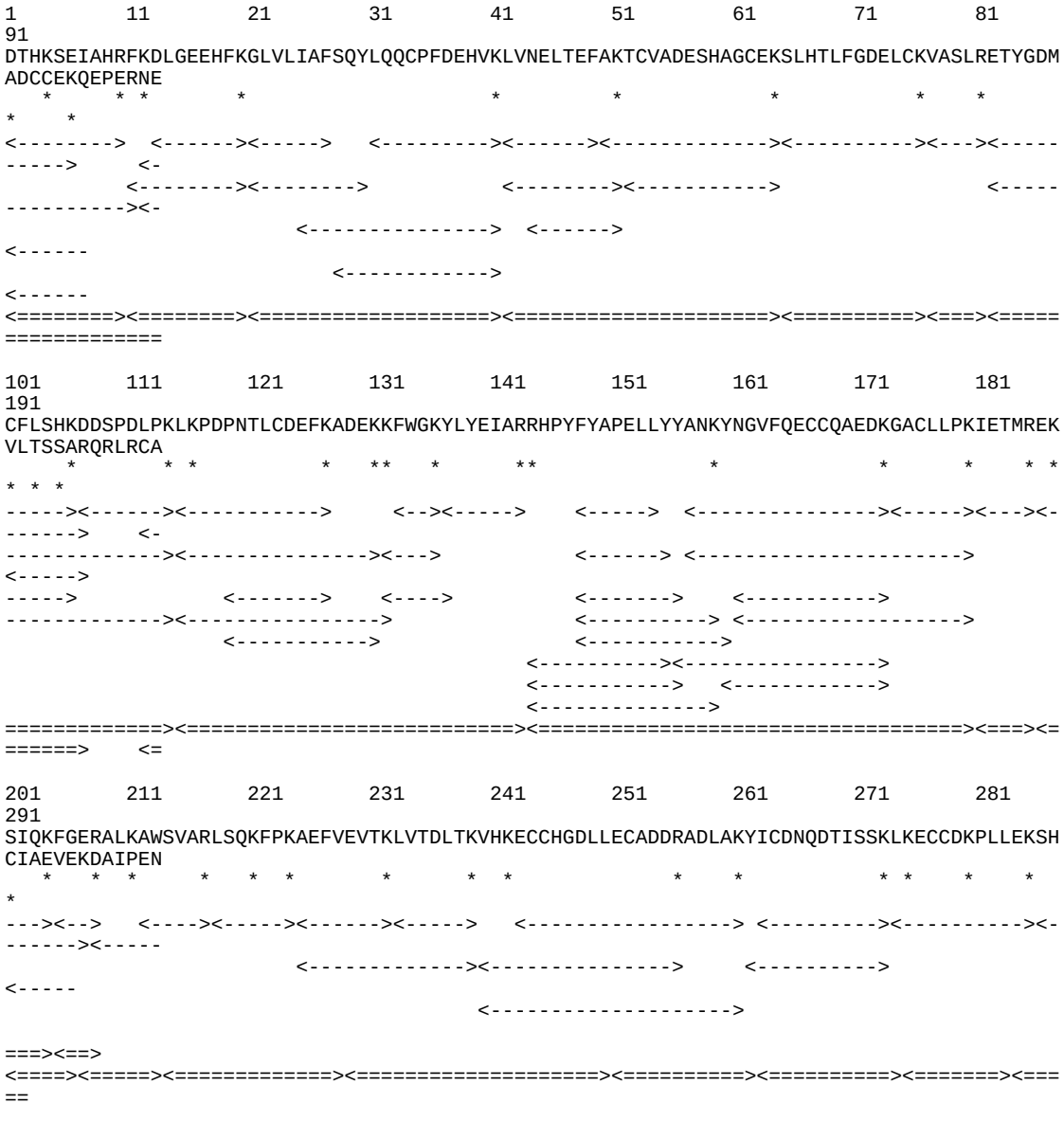


Figure B10. Tiling of the observed peptides with unique sequences on the mature sequence of BSA.

The coverage is calculated to be 558 amino acids out of a total of 583, which amount to 94.85%. The possible sites for trypsin cleavage, lysines and arginines are marked by asterisks.

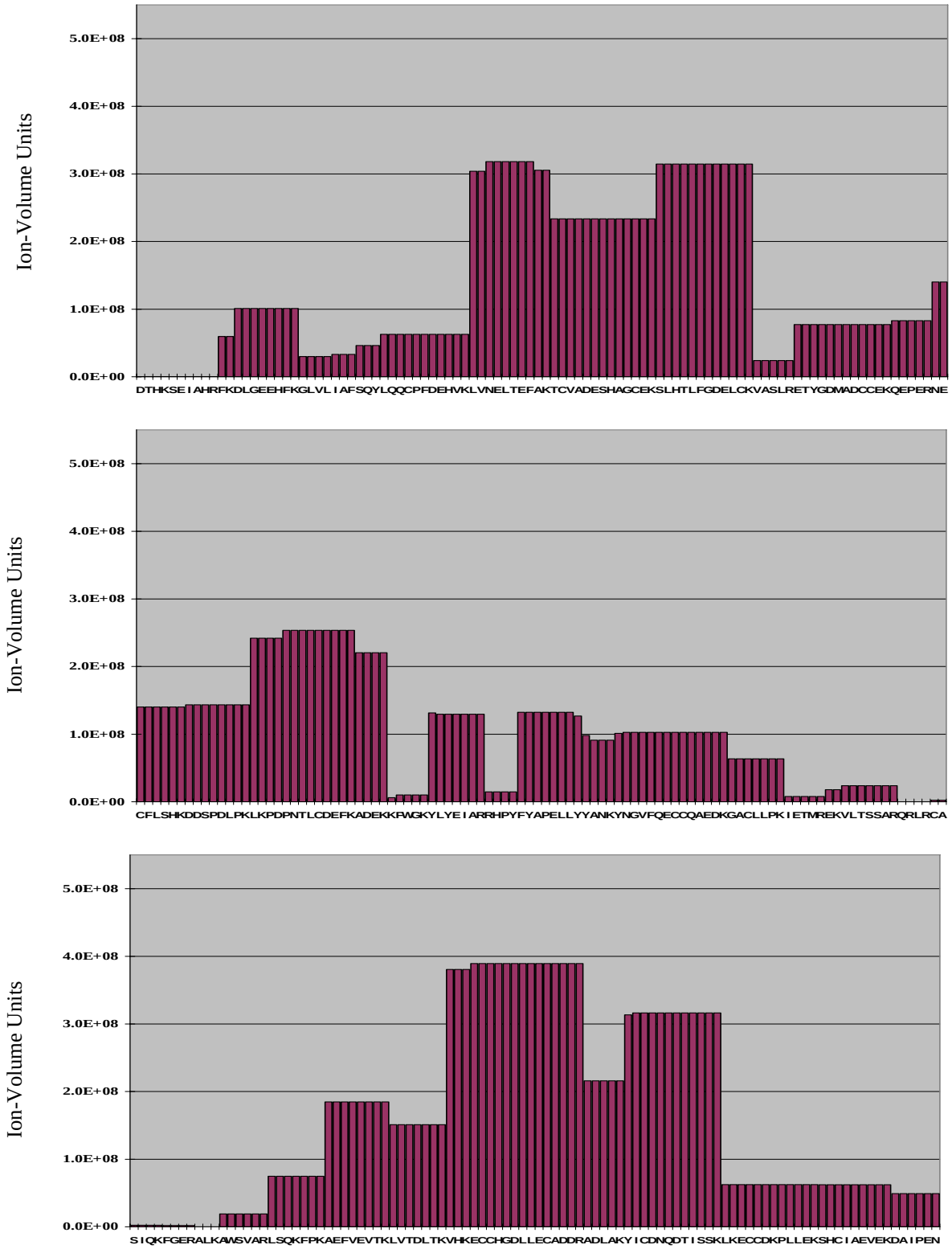


Figure B11. Cumulative ion volume for each amino acid in the protein (mature BSA).

The cumulative ion volume for each amino acid is calculated as the sum of the ion volumes for each peptide that contains that particular amino acid. The histograms show that certain areas of the protein are apparently more ionizable than others. However, this variation might also be due to unidentified peptides.

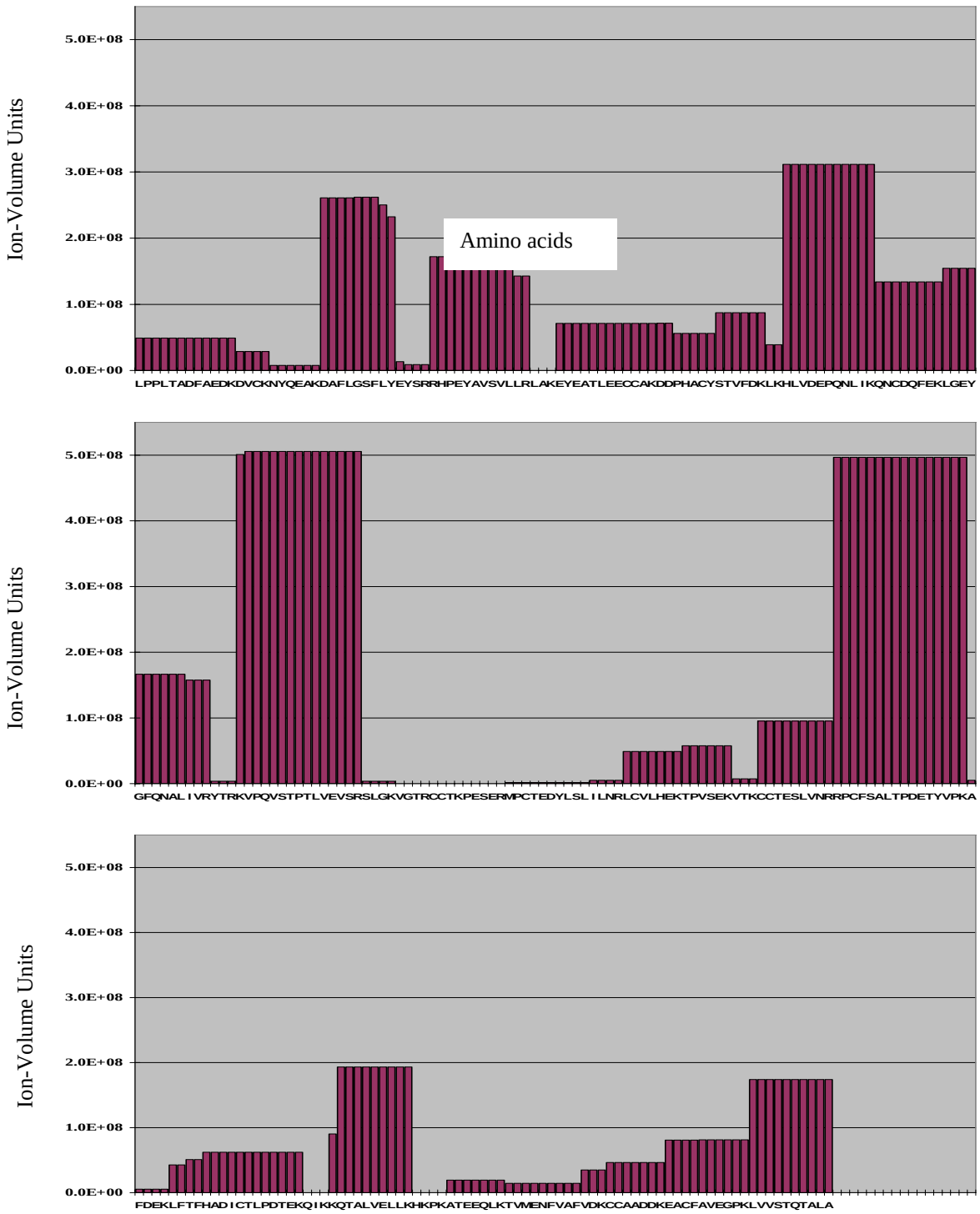


Figure B11 (continued)

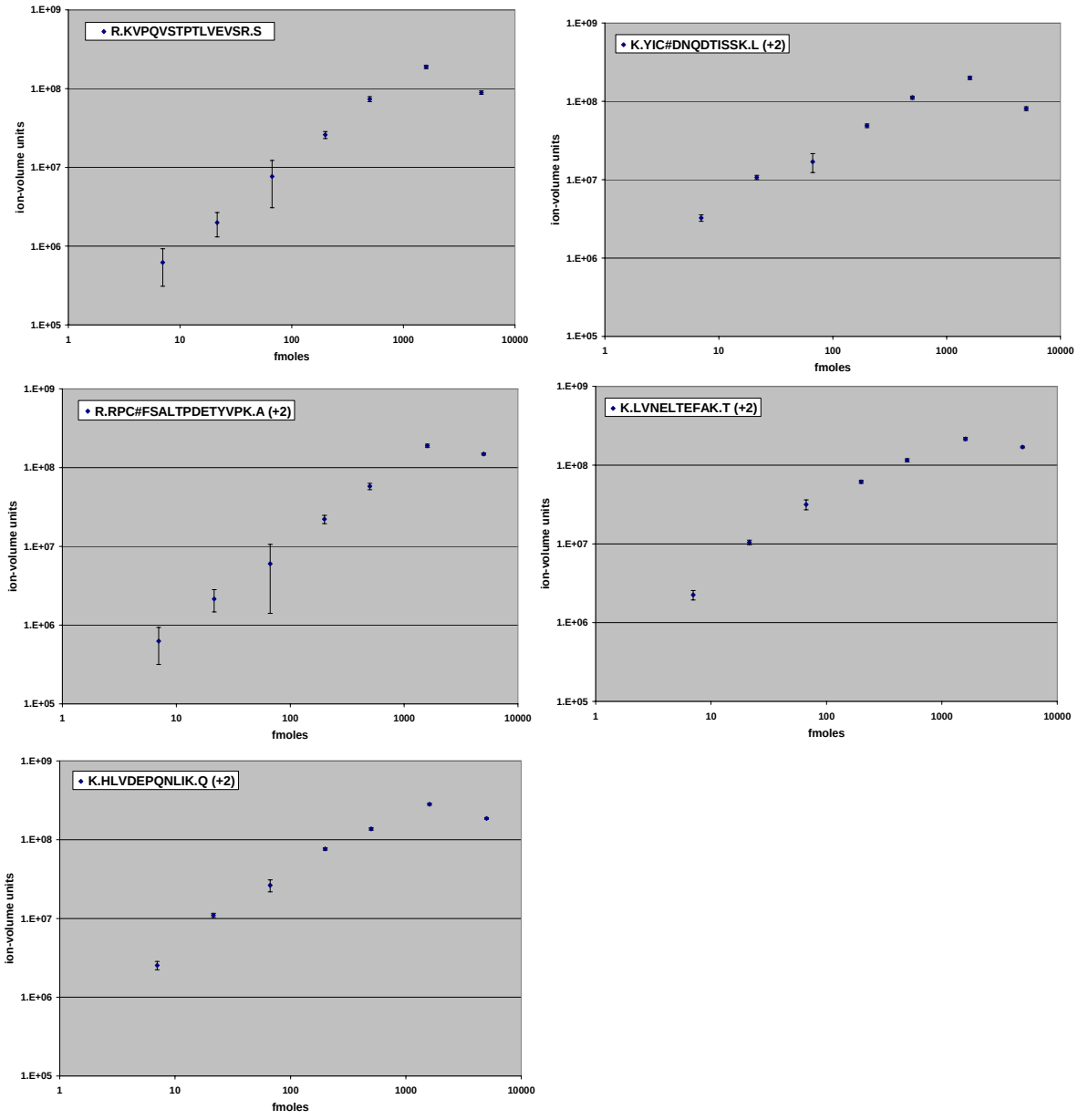


Figure B12. Calibration curves for the five most abundant +2 peptides.
The curves were generated using the abundances of 72 isotopic clusters, quantitated by MapQuant, either in a supervised (51) or an unsupervised manner (21).

5 Metabolic modeling Software

Identifying metabolic enzymes with multiple types of association evidence

Existing large-scale metabolic models of sequenced organisms commonly include enzymatic functions which can not be attributed to any gene in that organism. We present a novel method for identifying such missing genes based on a local structure of metabolic network and multiple types of functional association evidence, including clustering of genes on the chromosome, similarity of phylogenetic profiles, gene expression, protein fusion events and others. Using *E. coli* and *S. cerevisiae* metabolic networks, we illustrate the predictive ability of each individual type of association evidence and show that significantly better predictions can be obtained based on the combination of all data. In this way our method is able to predict 60% of enzyme-encoding genes of *E. coli* metabolism within the top 10 (out of 3551) candidates for their enzymatic function, and as a top candidate within 43% of the cases. Our approach does not rely on direct sequence homology to known enzyme-encoding genes, and can be used in conjunction with traditional homology-based metabolic reconstruction methods.

Supplementary materials are available at:
http://arep.med.harvard.edu/kharchenko/identification/supplements.html .

5.1. Motivation

Comprehensive and accurate reconstruction of the metabolic networks remains an important problem for both newly sequenced and well-studied organisms (Borodina et al. 2005; Reed et al. 2003). The challenges posed by the experimental determination of the metabolic enzymes have led to development of computational methods for metabolic reconstruction. The most common approach is to identify genes encoding a specific metabolic enzyme by establishing sequence homology to functionally characterized enzymes in other species (Tatusov et al. 1996). Although such sequence homology methods have been remarkably successful overall, they fail to identify enzymes encoded by genes with poor sequence homology to known metabolic enzymes, and result in partially reconstructed metabolic networks. The problem of identifying genes encoded for a specific metabolic function in such partially reconstructed networks has been referred to as the "missing gene" problem (Osterman and Overbeek 2003).

Computational strategies for identifying missing metabolic genes rely on refined sequence homology analysis (Green and Karp 2004; Reed et al. 2003) and consideration of functional association evidence linking candidate genes with known enzyme-encoding genes (Osterman and Overbeek 2003). For example, PathwayTools hole-filler developed by Green *et al*. (Green and Karp 2004), prioritizes candidates obtained from an initial sequence homology search by using, among other factors, information on whether the candidate gene is located adjacent to, or in the same transcriptional unit as known

enzyme-encoding genes of related metabolic function. In some cases, strong genome context association evidence, such as clustering of genes on the chromosome, or co-occurrence of genes in phylogenetic lineages, has played a key role in identifying metabolic genes in several organisms (Bishop et al. 2002; Bobik and Rasche 2001).

An extensive set of tools has been developed to detect and catalog general pair-wise functional associations between genes based on a combination of genome context methods and other evidence, such as co-expression or protein interactions (Bowers et al. 2004; von Mering et al. 2003). Combinations of heterogeneous association evidence have been used for general functional inference (Troyanskaya et al. 2003), prediction of protein complexes (Asthana et al. 2004; Jansen et al. 2003; Yamanishi et al. 2004) and synthetic lethal interactions (Wong et al. 2004). A recent work by Yamanishi *et al*. (Yamanishi et al. 2005) relied on a combination of genomic, mRNA expression and localization evidence, together with information on chemical compatibility to reconstruct metabolic pathways from known metabolic enzymes.

In an earlier study we described a method for identifying missing enzyme-encoding genes based on gene co-expression and local structure of metabolic network (Kharchenko et al. 2004). The candidate genes for encoding a missing metabolic enzyme were evaluated based on the overall similarity of their expression profile with the expression of the metabolic network neighborhood of the missing enzyme (Figure C1a). The local property of gene co-expression, which formed the basis of this method, was also observed for other types of functional associations, in particular for associations established by genome context (Kharchenko et al. 2005). In this work we showed that such an approach can be extended to identify metabolic enzyme-encoding genes from a number of different types of functional association evidence, including phylogenetic profile co-occurrence, physical clustering of genes on the chromosome and protein interaction data. We noted that the presented method does not rely on sequence homology to known enzymes, and its predictions are complementary to the traditional methods of metabolic reconstruction.

We illustrated the performance of each individual type of association evidence by testing how well the method is able to predict known enzyme-encoding genes of *E. coli* (Reed et al. 2003) and *S. cerevisiae* (Forster et al. 2003) metabolic models (see Methods). A set of candidate genes, containing all non-metabolic genes in an organism, is evaluated and prioritized by calculating overall association with the neighborhood of the missing metabolic enzyme (Figure C1b). To assess the performance of our method we relied on a self-rank measure, which is the rank of a known enzyme-encoding gene among the set of candidates prioritized for its own metabolic function (see Methods). We developed techniques for combining multiple types of association evidence and show that significantly better prediction performance can be achieved based on combined association evidence.

5.2. Results and discussion

Similarity of phylogenetic profiles

A number of earlier studies have explored using patterns of gene co-occurrence or absence in the phylogenetic lineages to infer functional association between gene pairs (Huynen and Bork 1998; Pellegrini et al. 1999). The basic premise of the method is that a function is likely to be encoded by several associated genes; therefore lineages maintaining only some of these genes will have lower evolutionary fitness. For instance, enzymes catalyzing successive steps of a linear metabolic pathway are likely to be present together in an organism relying on that metabolic pathway, and absent together from an organism that does not require that pathway.

A phylogenetic profile of a given gene on a set of N_G genomes can be encoded as binary string of length N_G, with each position marking presence (1) or absence (0) of an ortholog in the corresponding genome. Functional association between a pair of genes is assessed by the degree of similarity of their phylogenetic profiles. A number of different distance measures have been used to calculate such similarity, including Hamming string distance, mutual information and hypergeometric distribution (Bowers et al. 2004; Pellegrini et al. 1999, Huynen, 2000 #546; Wu et al. 2003). We find that the performance of different distance measures is very similar. These profile similarity measures do not take into account variable degree of divergence between genomes comprising the orthology dataset. This is particularly clear in the case of Hypergeometric distribution measure (Bowers et al. 2004; Wu et al. 2003), which assumes that ortholog occurrences are independently and identically distributed across the set of included genomes (see 5.3 Methods).

The identity assumption would suggest that the total number of ortholog occurrences within each genome should be approximately the same, and the distribution of the number of orthologs should form a single, narrow peak around an average ortholog number. The empirical distribution, however, is quite different from the expected form, lacking a peak around the mean, and showing substantial density over almost an entire range of ortholog numbers. When the identity assumption is relaxed, profile similarity probability is described by the Extended Multivariable Hypergeometric distribution (Harkness 1965). Because probability functions of this distribution have not been derived in a closed form, we developed a numerical algorithm for estimating these probabilities (see 5.3 Methods).

Bias stemming from the violation of the independence assumption can be minimized by exclusion or reduction of closely related species in the ortholog occurrence dataset. We developed an approach similar to previously published work (von Mering et al. 2003), which reduces the bias by folding together phylogenetic branches containing closely related species, and using an ortholog occurrence pattern based on the agreement within the folded branch (see 5.3 Methods).

The effect of both corrections on the ability to predict enzyme-encoding genes in *E. coli* is illustrated by the cumulative self-rank distributions (see 5.3 Methods) in Figure C2a. The extended hypergeometric distribution correction for the variable genome divergence from *E. coli* target genome (violation of the identity assumption) provides a noticeable improvement in prediction performance (8% at self-rank threshold of 50). On the other hand, the folding method correcting for variable divergence with the set of query genomes (violation of independence assumption) does not significantly improve the results.

The phylogenetic profile co-occurrence method depends on identification of orthologous genes across potentially diverse lineages. Existing investigations have used a variety of methods, including readily available *Clusters of Orthologous Groups* (COG) database (Tatusov et al. 2001), (von Mering et al. 2003; Wu et al. 2003), closest homologs (Bowers et al. 2004), and best bi-directional homology pairs (Huynen et al. 2000). The results presented in our work rely on two alternative sets of orthology data. The first set comes from KEGG SSDB database (Itoh et al. 2004), and includes closest homologs and best bi-directional hits as determined by the Smith and Waterman algorithm (we will refer to it as KEGG-based dataset). The second set was constructed based on results of BLAST (Altschul et al. 1990) queries against a "non-redundant" set of known protein sequences maintained by NCBI (see Methods). The set also includes information on reverse BLAST searches to determine best bi-directional hits (referred to as BLAST-based dataset).

Predictive performance of different orthology datasets is compared in Figure C2b. We note that coverage of the COG orthology data is biased towards genes encoding known metabolic enzymes, and the self-rank performance of this dataset was estimated by normalizing with respect to the non-metabolic gene coverage. Figure C2b shows that profile associations calculated using BLAST-based dataset provide better predictions of enzyme-encoding genes than association based on the KEGG orthology dataset. We also find that in the case of both datasets better performance is attained when using best bi-directional homology pairs instead of closest homologs.

As a consequence of gene duplications, metabolism contains a significant number of paralogous enzyme pairs (Maltsev et al. 2005). In many cases, such enzymes continue to catalyze the same reactions. Such pairs will frequently have similar or identical orthology mappings, and their inclusion can lead to a significant bias in estimation of the predictive performance. The results presented in this work, therefore, exclude self-ranks of any metabolic enzymes that have high sequence homology to any other metabolic enzyme in the organism (see 5.3 Methods).

Co-expression of orthologous genes

The approach for identifying enzyme-encoding genes based on the similarity of mRNA expression profiles (Kharchenko et al. 2004) can be extended to include co-expression information of orthologous genes in other organisms. Conservation of mRNA co-expression across different species has been investigated by a number of recent studies

(Bergmann et al. 2004; Snel et al. 2004; Teichmann and Babu 2002; van Noort et al. 2003). For example, analysis of co-expressed gene pairs between *S. cerevisiae* and *C. elegans* shows statistically significant (P value < 10^{-3}) level of conservation (van Noort et al. 2003). Although the number of pairs with highly conserved co-expression is small, incorporating ortholog co-expression can provide significant improvements to the accuracy of functional predictions based on the mRNA expression data (Teichmann and Babu 2002; van Noort et al. 2003).

We find that enzyme-encoding gene predictions based on the co-expression of *E. coli* orthologs in *S. cerevisiae* achieve good performance on the enzymes covered by the dataset. Although *S. cerevisiae* orthologs can be identified for only 40.1% of *E. coli* metabolic genes, combining native and ortholog co-expression scores provides noticeable improvements. Combination of native and ortholog co-expression increases the fraction of metabolic enzymes predicted within the top 50 candidates from 27% to 36%. Similarly, using *E. coli* expression data improves prediction results for enzyme-encoding genes of *S. cerevisiae* metabolic network. Overall self-rank performance based on combined co-expression data is included in Figure C4.

Clustering of genes on the chromosome

Relative positions of genes on the chromosome have also been successfully used to infer functional associations. Most notably, analysis of prokaryotic genomes focused on identifying pairs of orthologs located close to each other on the chromosome, as well as sets of such pairs (Overbeek et al. 1999; von Mering et al. 2003; Yanai et al. 2002). Such clustering is also observed in the eukaryotic genomes, even though they lack well-defined operon structures. A recent study by Lee *et al.* (Lee and Sonnhammer 2003) analyzed clustering of genes in KEGG pathways for 5 distant eukaryotic species. The study demonstrated that depending on the genome, 30% to 98% of the pathways exhibit statistically significant levels of gene clustering on the chromosome. A variety of methods have been developed for identifying chromosome gene clusters and evaluating their significance (Durand and Sankoff 2003). To generate association scores we use a simple statistical evaluation strategy based on the chromosome gene order, which allows for computationally efficient treatment of large number of genomes (see Methods).

The self-rank performance based on the chromosome clustering association is shown in Figure C4. The overall performance for known *E. coli* metabolic enzymes is better than for the *S. cerevisiae* enzymes, which is expected given the prominent role of operons in prokaryotic transcriptional regulation.

Other association measures

Interacting proteins encoded by separate genes in some species, may sometimes occur as a single, multi-domain fusion protein in other species. Detecting fusion of non-homologous proteins in another organism has been shown to be a significant predictor of functional association between genes (Enright et al. 1999; Marcotte et al. 1999; Yanai et al. 2001). Our calculations of a fusion association score are based on a combination of fusions detected at several sequence homology thresholds. The overall performance of the method is included in Figure C4. Although protein fusion associations are only able

to predict relatively small fraction of enzyme-encoding genes (18% for *E. coli*), almost all of predicted enzymes are returned within the top 20 candidates.

A number of metabolic reactions are catalyzed by well established protein complexes, such as the phosphofructokinase complex. Furthermore, metabolic processes commonly involve interactions between multiple metabolic enzymes. For instance, the phosphofructokinase alpha subunit encoded by Pfk1 also interacts with a product of Fba1, fructose-biphosphate adolase II, catalyzing an adjacent reaction in the glycolysis pathway (Matic et al. 2001). Large protein-protein interaction datasets have been generated by studies using yeast two-hybrid systems (Ito et al. 2000; Uetz et al. 2000) and, more recently, mass spectrometry-based techniques (Gavin et al. 2002; Ho et al. 2002). In the framework of our approach, candidate genes can be evaluated by assessing the overall amount of interactions between a candidate gene and the metabolic network neighborhood of a missing enzyme. To assess confidence of individual interactions, our analysis makes use of the probabilistic protein interaction dataset from Jansen *et al*. (Jansen et al. 2003), which combines results of four high-throughput interaction datasets (Gavin et al. 2002; Ho et al. 2002; Ito et al. 2000; Uetz et al. 2000) . The performance of our prediction method on the protein interaction data is significantly lower than that of other association scores, nevertheless it is above what is expected from a random association score (Figure C4b).

Functional association can also be assessed through similarity of deletion mutant phenotypes under a large set of environmental conditions. For example, deletions of genes that are adjacent to each other in a linear metabolic pathway are likely to result in identical mutant phenotypes. A recent work by Dudley *et al.* (Dudley et al. 2005) experimentally measured growth phenotypes of 4710 *S. cerevisiae* mutants under 21 experimental conditions, including different carbon sources, nutrient limitations, stress and others conditions. We tested the performance of our prediction algorithm on a set of 53 known metabolic enzyme-encoding genes for which high-confidence data was available. While the results illustrate predictive power of phenotypic profile associations, overall contribution of this score to the predictions of unidentified enzyme-encoding genes is very small. This is expected, because available high-confidence phenotypic data covers only 14% of *S. cerevisiae* genes.

Overall enhancements of the individual association scores

Description of a metabolic network neighborhood can be enhanced by considering relative strength of metabolic connections established by different metabolites. Metabolites connecting many enzyme-encoding genes pairs establish, on average, weaker functional associations (Kharchenko et al. 2005). The performance of our predictive method can be improved by weighting the contribution of each neighbor in evaluating the overall association of a candidate gene with the metabolic network neighborhood of a missing enzyme. The weight is assigned according to the total number of enzyme pairs associated with a connecting metabolite (see 5.3 Methods).

Distributions of association scores between a given gene and all other genes in an organism tend to differ from one gene to another. For instance, a gene whose orthologs

can be identified in many organisms will typically have more high-confidence chromosome clustering associations than a gene with relatively few detected orthologs. This introduces bias when evaluating overall association with a metabolic network neighborhood. The association-rank rescaling reduces this bias by translating raw association scores into probabilities of metabolic adjacency, calculated based on the rank of raw association score within a distribution of all scores for a particular gene. The rescaling procedure also reduces the number of false positives by considering raw association score of a gene pair with respect to organism-wide score distributions of both genes and choosing a more conservative adjacency probability value.

The predictive performance of all association scores is improved by either correction, with the exception of the protein fusion score, where application of metabolite weighting results in weaker performance.

Predictions based on combined association evidence

Enzyme-encoding gene predictions based on the individual association scores can be combined to achieve better performance. Normalizing relative strength of different association scores requires informative priors. Such priors can be either constructed manually, for example by consulting experts (Troyanskaya et al. 2003), or learned from known test-cases. This problem has been extensively considered with respect to confidence in pair-wise gene functional associations, and test cases for learning the priors were based on known functional groupings, such as GO annotations (Lee et al. 2004) or membership in KEGG pathways (von Mering et al. 2003). For the current problem of prioritizing enzyme-encoding gene candidates, such priors can be learned from known enzyme-encoding genes (Green and Karp 2004).

Towards the goal of integrating multiple types of association evidence, we have developed two distinct methods. The first approach is based on a *direct likelihood-ratio* (DLR) evaluation of the association score probability distributions. The likelihood that a given candidate gene encodes the desired metabolic enzyme is calculated under the simplifying assumptions that individual association scores are independent and monotonic. The monotonic assumption states that for every association score, the likelihood of association increases monotonically with the absolute value of the score. Both assumptions allow for useful approximations, but in general can be shown to be incorrect. For example, clustering of genes on the chromosomes in *E. coli* is statistically significantly correlated with the similarity in expression profiles (Spearman rank correlation P value $< 10^{-10}$), violating the independence assumption. The DLR method calculates overall likelihood ratio of a candidate gene encoding the desired enzyme as a product of likelihood ratios for each individual association score (see Methods).

The second approach uses a general machine learning method called Adaboost (Freund and Schapire 1997; Schapire 2002), and does not rely on independence or monotonicity of the association scores. The generated classifiers are in the form of *alternating decision trees* (ADT), which are generalization of decision stumps, decision trees, and their combination (Freund and Mason 1999). In addition to flexible semantic representation, ADT-based classifiers provide a real-valued measure of confidence, called *classification*

*marg*in, which can be related to the probability of a given classification being correct (Schapire et al. 1997). The Adaboost method has been successfully applied to several large-scale biological problems, including detection of transcription factor binding motifs and prediction of regulatory response (Middendorf et al. 2005; Middendorf et al. 2004).

We find that in identifying missing metabolic genes both ADT and DLR methods achieve comparable levels of performance (Figure C3). The ADT method performs slightly better on *E. coli* metabolic enzymes and DLR on *S. cerevisiae*. Success of the DLR method relative to a general classifier, such as ADT, suggests that the derived association scores are largely consistent with the underlying assumptions of monotonicity and independence, and allow quality predictions to be made based on a straightforward evaluation of the score probability distributions. The ADT method, however, does not require such assumptions, and may be used to incorporate in the future a wide variety of unrestricted descriptors, such as sequence homology data or expression variability (Kharchenko et al. 2004).

Prediction performance of individual functional association scores and their combination using ADT method is shown for *E. coli* metabolic enzymes in Figure 4a. The figure illustrates that predictions based on the combined evidence are clearly superior to what is achieved by any individual type of functional association evidence, with 43% of known enzymes predicted as number one candidates for their enzymatic function, and 60% within the top 10 candidates. Associations based on the chromosome clustering provide the best predictions of any single evidence type, and are able to predict almost half of the metabolic enzymes within the top 10 candidates. It is also important to note that different association evidence types are not redundant – none of the predictions based on a particular association score are completely covered by the predictions of another association score.

Individual and combined prediction performance for enzymes of *S. cerevisiae* metabolic network is illustrated in Figure 4b. Relative to *E. coli* predictions, co-expression score in *S. cerevisiae* tends to perform better; however chromosome clustering and phylogenetic profile association scores perform worse. The overall level of performance is also lower, with approximately 60% of the enzymes predicted within top 50 candidates (compared to 71% in *E. coli*). The performance difference can be partially attributed to lower number of candidate genes in *E. coli* (3351 as opposed to 5252 in *S. cerevisiae*) and wider availability of the genomic data for bacterial organisms. For example, chromosome clustering associations were calculated on a dataset that contains nearly a hundred bacterial species and only a handful of eukaryotic genomes.

5.3. Methods, Assumptions, and Procedures

Metabolic neighborhoods and network representation

The metabolic network was represented as a graph, with nodes corresponding to metabolic enzyme-encoding genes and edges to connections established by the metabolic reactions (Kharchenko et al. 2005). Two metabolic genes are connected if the enzymes they encode share a metabolite among the set of reactants or products of the reactions they catalyze. Metabolic *network distance* between enzyme-encoding genes is calculated as a shortest path in the graph. Distance of directly connected genes is taken to be 1. A *metabolic neighborhood layer* of a radius R around a metabolic enzyme X is defined as a set of all enzyme-encoding genes that are at the distance R from the enzyme X. A *metabolic neighborhood* of radius R is a set of neighborhood layers of radii $r \leq R$ (Figure C1a).

Detailed metabolic models of *E. coli* (Reed et al. 2003) and *S. cerevisiae* (Forster et al. 2003) were used to compile comprehensive connectivity graphs for these organisms, excluding metabolic connections established by the following top 14 most common metabolites: ATP, ADP, AMP, CO_2, CoA, glutamate, H, NAD, NADH, NADP, NADPH, NH_3, orthophosphate and pyrophosphate (and corresponding mitochondrial and external species).

Self-rank validation

To assess performance of our method we use self-rank measure, which quantifies the ability to predict known metabolic enzymes. A self-rank of a known enzyme-encoding gene is defined as a rank of that gene among a set of candidates in an ordering determined by our algorithm (Figure C1b). A set of candidates consist of all genes in the organism that do not already appear in the metabolic graph (i.e. non-metabolic genes) and the known enzyme-encoding gene that is being tested. A candidate set for *E. coli* contained 3351 open reading frames (ORFs), and for *S. cerevisiae* 5252 ORFs. A perfect prediction algorithm would result in a self-rank of 1 (top candidate) for every metabolic enzyme, and a completely non-informative method would result in a uniform distribution of ranks (on the range from 1 to the size of the candidate set).

The overall performance of the method was measured by evaluating self-ranks of a set of known enzyme-encoding genes. This set contains all known metabolic enzymes in an organism, except for the enzymes that have high sequence homology (BLASTp E value below 10^{-10}) to some other known metabolic enzyme in that organism (paralogs). The exclusion of such paralogous pairs aims to avoid bias stemming from overlapping ortholog mappings. The resulting set contained 351 enzymes from *E. coli* metabolism, and 240 from *S. cerevisiae*.

Orthology datasets

The KEGG ortholog dataset was retrieved from Sequence Similarity Data

Base (SSDB) (01/2005). All available closest homologs and best bi-directional hits of *E. coli* and *S. cerevisiae* genes were recorded. The BLAST-based dataset was constructed using BLASTp queries against NCBI NR protein dataset (03/2005), using E-value cutoff of 10^{-3} and limiting the maximum number of homologs per query to 6000. To determine best bi-directional hits, reverse BLASTp queries were run for every hit against target genome (*E. coli* or *S. cerevisiae*). NCBI taxonomy identifiers were used to group hits belonging to the same organism. For *E. coli* only organisms containing orthologs to more than 4% of genes were considered (7% for *S. cerevisiae*). We found that performance of analogous datasets constructed using TBLASTN queries was similar.

Phylogenetic profile co-occurrence

Given a set of genomes $G = \{G_1..G_{N_G}\}$, a phylogenetic profile of a gene was represented as a binary vector p of length N_G, such that $p_i = 1$ if an orthologous gene is present in genome G_i, and $p_i = 0$ otherwise.

Assuming that orthologs are independently identically distributed (IID) *with*in each genome G_i, the probability of observing two profiles of a given similarity under the null hypothesis is calculated using hypergeometric distribution (Wu et al. 2003):

$$P(k|n,m,N) = \frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}}$$

where k is the number of ortholog co-occurrences, N is the size of the genome set G, n and m correspond to the number of orthologs in the two profiles being compared. The probability of functional association is then given by $P_{association} = 1 - \sum_{k>K} P(k|n,m,N)$, where K is the number of actual ortholog co-occurrences observed between two specific profiles (Bowers et al. 2004).

If the assumption of identical ortholog distribution within each genome is relaxed, probability $P(k|n,m,N)$ is distributed as a sum of independent, non-identical Bernoulli variables x_i: $k \sim \sum_{\min(n,m)} x_i$, with $p(x_i)$ corresponding to the probability of observing a match in a given genome i. This is a special case of the Extended Multivariable Hypergeometric distribution (Harkness 1965).

To determine the probability of observing k ortholog co-occurrences between profiles of a given gene x and some other gene y, $P(k|n,m,N)$, we calculate $P_N(k|n,m)$ given by the recursive formula below. In general, $P_i(k'|n',m')$ is the probability of observing k' ortholog co-occurrences at a current (G_i) or subsequent genomes as we walk along a

predefined genome order, where n' and m' is the number of orthologs of gene x and gene y respectively in genomes G_j such that $j \le i$. $P_i(k'|n',m')$ is defined recursively as: $P_i(k'|n',m') = p^o_{i,m'}\left[p^c_i P_{i-1}(k'-1|n'-1,m'-1) + (1-p^c_i)P_{i-1}(k'|n',m'-1)\right] + (1-p^o_{i,m'})P_{i-1}(k'|n',m')$ where i is the genome index, $p^o_{i,m'}$ is the probability of one of the remaining m' orthologs of gene y occurring in genome G_i, and p^c_i is the probability of ortholog co-occurrence in the genome G_i.

Given the values of probabilities $p^o_{i,m'}$ and p^c_i, the value of $P_i(k'|n',m')$ is computed using a dynamic programming approach. p^c_i is equal to 0 or 1, depending on whether an ortholog of gene x is present in genome G_i. The consideration of non-identical distribution of ortholog frequency within each genome is then localized to $p^o_{i,m'}$, which in this case is distributed according to the marginal Extended Hypergeometric distribution. The marginal form of the distribution is more amenable to the computational approximations than the regular form. Since $p^o_{i,m'}$ does not depend on the choice of genes x and y, we sample $p^o_{i,m'}$ computationally, taking into account individual ortholog occurrence frequencies of each genome. The probability of ortholog occurrence in a specific genome ($p^o_{i,m'}$) was sampled computationally by drawing from the set of organisms without replacement with relative probabilities corresponding to the rate of ortholog occurrences in each genome. In each iteration, draws were performed until all of the organisms were drawn. A total of 10^6 such iterations were performed.

To correct for non-independent ortholog occurrence rates, we first evaluate the distance between a pair of query genomes X and Y as $d(X,Y) = \frac{MI(X,Y)}{\min(H(X),H(Y))}$, where $MI(X,Y)$ is mutual information between ortholog occurrence vectors for genomes X and Y, and $H()$ is Shannon entropy of each vector. The ortholog occurrence vector for a query genome *X* is a binary vector of length N_{genes} (number of genes in a target organism, i.e. *E. coli*), such that the value of the i^{th} element is 1 if ortholog of an i^{th} gene is found in *X*, and 0 otherwise. Clusters of closely related organisms ($d < 0.8$) were determined by the greedy neighbor-joining method. Several ways of summarizing the ortholog co-occurrence vector for a cluster of closely related organisms were tested: selecting the organism with highest entropy, using AND/OR functions, and using majority rule. We find that performance of AND function is optimal for the threshold of $d < 0.8$, however for higher thresholds selecting an organism with highest entropy results in better performance.

In evaluating performance without adjacency-rank rescaling (Figure C2), total phylogenetic profile association score between a candidate gene x and a metabolic neighborhood layer L was calculated as $score_L(x) = \sum_{g \in L} \frac{1}{P(x,g)}$, where $P()$ is the

probability of observing a given number of ortholog co-occurrences calculated using one of described probability distributions.

To estimate self-rank performance of the COG dataset correcting for the bias in orthology dataset coverage (Figure C2b), the fraction of true enzyme-encoding genes, f predicted within a particular self-rank threshold t was calculated as $f(t) = \alpha_M f'(\alpha_C t)$, where α_M is the fraction of test metabolic enzyme-encoding genes covered by the COG dataset, α_C is the fraction of candidate set genes (non-metabolic) covered by the dataset, and f' is the performance on the set of metabolic and candidate genes covered by the COG dataset.

Gene co-expression

The co-expression association value was calculated as a Spearman rank correlation (Press et al. 2002) between expression profiles. *E. coli* co-expression was calculated based on the 180 conditions from the Stanford Microarray Database (Sherlock et al. 2001). *S. cerevisiae* co-expression was measured based on the mRNA expression profiles from Rosetta "compendium" dataset (Hughes et al. 2000). Log10 intensity ratio data was used. Co-expression of orthologous genes was determined using the KEGG ortholog dataset.

Clustering on the chromosome

The degree to which orthologs of two genes are clustered on the chromosome was calculated based on the null hypothesis that genes are randomly distributed across the chromosomes. Instead of considering gene sizes and exact nucleotide positions, we concentrated on gene order statistics. The association strength was determined as the probability of chromosome gene order position of a candidate gene *x* for a metabolic neighborhood layer N_l: $P(x \mid N_l) = \prod_{y \in N_l} \prod_{g \in G} P\left(d_g(x,y)\right)$, where G is a set of query genomes in which orthologs of both *x* and *y* can be found, and $P\left(d_g(x,y)\right)$ is the probability of observing gene order distance $d_g(x,y)$ between genes *x* and *y* in a genome *g*. This was calculated directly, based on the organism chromosome sizes under the null hypothesis. The above formulation is based on two major assumptions: (1) gene order distances to different genes of the neighborhood layer N_l are independent, and (2) gene order distances between a specific pair of genes are independent across different organisms.

The results are based on a set of 105 bacterial and three eukaryotic genomes (*S. cerevisiae, S. pombe, C. elegans*) from Genbank. The set was screened to eliminate closely related species using ortholog occurrence mutual information threshold of 0.9. Orthology mapping was established using KEGG SSDB best bi-directional hits.

Protein interactions

Interaction likelihood ratios from the PIE dataset by Jansen *et al*. (Jansen et al. 2003) were used as pair-wise protein interaction association values.

Protein fusions

Two proteins *x* and *y* of a target genome (*S. cerevisiae* and *E. coli*) were taken to be associated through a protein fusion event if both of the following conditions were met:

1.) *x* and *y* are homologous to the same protein *z* in one of the query genomes with a BLASTp E value below a specified threshold ($E_{threshold}$), and with at least 70% of their sequences aligned to *z*.
2.) *x* and *y* align to different regions of *z*, or to regions overlapped by no more than 10% of the shorter protein among *x* and *y*. If *x* or *y* align to multiple regions of *z*, then any two regions must not overlap.

A set of 70 query genomes, based on the study by Bowers *et al*. (Bowers et al. 2004), was downloaded from the Entrez Genome database: (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome).

Several values of $E_{threshold}$ were used in generating enzyme-encoding gene predictions (Figures C3 and C4), with $E_{threshold}$ =10^{-2}, 10^{-5} and 10^{-10} for *E. coli*; $E_{threshold}$ =10^{-3}, 10^{-5} and 10^{-10} for *S. cerevisiae*.

Adjacency-rank score rescaling, metabolite weighting and calculation of layer association scores

To perform adjacency-rank score rescaling of raw pair-wise association values, we calculate the adjacency likelihood ratio for a pair of genes *x* and *y* as: $alr(x,y) = \frac{N_{genes}}{\max\{r_x^y, r_y^x\}} P_{adj}\left(r \le \max\{r_x^y, r_y^x\}\right)$, where r_y^x is a rank of gene y among a set of raw association values between gene x and all other genes in the organism. Lower ranks correspond to higher stringency of association. The probability P_{adj} is calculated from an empirical distribution of all ranks r_a^b, such that genes a and b are adjacent to each other (i.e. directly connected) in the metabolic network. N_{genes} is the number of genes in an organism.

Without metabolite weighting, the total association score between a candidate gene x and a metabolic neighborhood layer L is calculated as: $score_L(x) = \sum_{g \in L} \exp[alr(x,g)]$.

Metabolite weighting is incorporated by calculating total association score as: $score_L(x) = \sum_{g \in L} w_g \exp[alr(x,g)]$, where $w_g = \prod_{m_i \in \Theta} \frac{1}{N_{pairs}^{m_i}}$, m_i is the i^{th} metabolite in the shortest path Θ connecting neighborhood gene g with the missing enzyme. $N_{pairs}^{m_i}$ is the total number of gene pairs connected by a metabolite m_i. If more than one metabolite connects genes along the path Θ, a metabolite with the smallest N_{pairs} is used.

Direct likelihood-ratio predictor method

The placement algorithm considers each candidate gene by evaluating $P(M \mid D)$, which is the conditional probability that a given candidate encodes the desired enzyme (model, M) given all available evidence (data, D). Following Bayes rule we can

calculate that probability (up to a constant) using: $P(M \mid D) \propto \frac{P(D \mid M)}{P(D)}$, where $P(D \mid M)$ is the probability of observing existing associating evidence for a true enzyme-encoding gene. Assuming that different types of associative evidence scores are independent, we calculate probabilities as: $P(D) = \prod_e P_e(D_e)$, where $P_e(D_e)$ corresponds to the posterior of evidence type e. The problem is therefore transformed into estimating tails of association score probability distributions over all genes, and enzyme-encoding genes. For different types of associating evidence scores these probabilities were evaluated empirically from the gene counts, assuming that the likelihood of association increases monotonically with the absolute value of the score.

The self-rank evaluations of known *E. coli* and *S. cerevisiae* metabolic enzyme-encoding genes (see Self-rank validation section in 5.3 Methods) were performed using a leave-one-out validation strategy. In other words, in each case, scores of the candidate being evaluated are not included when calculating $P(M \mid D)$.

Alternating decision tree predictor

The *mljava* implementation of the *AdaBoost* algorithm (Freund and Mason 1999) was used to build ADT classifiers based on a set of descriptors, corresponding to different association scores with individual layers of the metabolic network neighborhood. The results presented in Figures C3 and C4 are based on 10-fold validation, 100 iterations of boosting. The training sets included data on only 60% of the true negative (non-metabolic) genes in order to minimize computational time. The candidate genes were prioritized according to the value of the classification margin.

Predictions with combined association evidence

The self-rank performance illustrated in Figures C3 and C4 was calculated based on candidate association with first three layers of metabolic network neighborhood. Association with respect to each layer was described by a separate association score. The predictions were performed using association score ranks: given a candidate gene x for a missing enzyme e, the value of a descriptor was calculated as a rank of $score_L(x)$ in a set of scores $S = \{score_L(y)\}_{y \in C}$, where C is a set of all candidates for a missing enzyme e, with higher ranks corresponding to stronger associations. For the *E. coli* metabolic model, the following association scores were used:

- Phylogenetic profile co-occurrence, calculated with extended hypergeometric and folding corrections, on orthologs established by best bi-directional homology relationship. Separate scores were calculated using BLAST-based and KEGG-based orthology data.
- Chromosome clustering.
- Gene co-expression. Separate scores were calculated for *E. coli* expression data, and for expression of *E. coli* orthologs in *S. cerevisiae* dataset.
- Protein fusion. Separate scores were calculated for different values of $E_{threshold}$: 10^{-2}, 10^{-5} and 10^{-10}.

Analogous scores were used for predictions on *S. cerevisiae* metabolic model, with addition of a protein interaction score. In the case of protein fusion, the scores were calculated for the following values of $E_{threshold}$: 10^{-3}, 10^{-5} and 10^{-10}.

5.4. Conclusion

The results presented in this work demonstrate that the gene encoding a specific metabolic function can be effectively identified from combined functional association with the metabolic network neighborhood of the desired function. This indicates that the relationships established by the local structure of the metabolic network impose constraints on a wide range of natural processes, such as gene expression or evolutionary processes on both molecular and genomic scales. Our tests used a combination of genome context and expression data to identify known *E. coli* metabolic enzymes, predicting them within the top 10 (out of 3351) candidates in 60% of the cases. We show that in the case of both *E. coli* and *S. cerevisiae*, combining multiple types of association evidence results in a significantly better prediction performance than that of any individual association score.

In validating the performance of our method, we relied on the metabolic network neighborhood as the sole source of information about the desired enzymatic activity. In practice, additional clues regarding activity or physical properties of the unidentified enzyme can often be used to narrow down the set of candidates. These additional clues may provide restrictions on the phylogenetic profile pattern, protein size, presence or absence of membrane spanning regions or specific protein domains. For example, for *E. coli* arabinose-5-phosphate isomerase, *yrbH* (Meredith and Woodard 2003) is predicted as a 10^{th} candidate among all genes, but is the only candidate within the top 50 with a putative sugar isomerase domain.

Sequence homology to known proteins remains the primary method of identifying missing enzymes (Huynen et al. 2003; Osterman and Overbeek 2003). Predictions based on the association evidence considered in this work are complementary to homology-based methods, and can be used to target enzymes that have not been identified in any organism (referred to as *globally miss*ing *enzymes* by Osterman *et al*.). Integration of genome context information into the refined sequence homology searches has been shown to improve the predictions (Green and Karp 2004). It will be important to analyze how incorporation of diverse association evidence presented in this work would improve the performance, in particular with respect to the difficult cases of weak or ambiguous sequence homology. The overall performance of the presented method can be improved in a number of ways. The datasets underlying individual scores can be expanded. Genome divergence corrections for the chromosome clustering score are also likely to improve the results. Further extensions can provide better identification in the cases where multiple missing genes appear within the same metabolic neighborhood. We hope that the
presented method, and its future derivations, will be important in completing metabolic models of different organisms.

Table C1: Association scores used in self-rank tests on combined evidence

Table C1: Association scores used in self-rank tests on combined evidence

Evidence type \ Organism	*E.coli*	*S.cerevisiae*
Phylogenetic profile co-occurrence.	• BLAST-based dataset score • KEGG-based dataset score Pairwise associations were calculated using extended hypergeometric and folding corrections, on orthologs established by best bi-directional homology relationship.	
Clustering of genes on the chromosome	• Gene clustering scores. Pairwise associations were calculated on 108 genomes, with KEGG-based orthology dataset	
Gene co-expression	• *E. coli* SMD expression dataset score • Expression of *E. coli* orthologs in *S. cerevisiae* Rosetta dataset.	• *S. cerevisiae* Rosetta expression dataset score • Expression of *S. cerevisiae* orthologs in *E. coli* SMD dataset.
Protein fusion	Separate scores were calculated for different values of $E_{threshold}$: • 10^{-2} • 10^{-5} • 10^{-10}	Separate scores were calculated for different values of $E_{threshold}$: • 10^{-3} • 10^{-5} • 10^{-10}
Protein interactions		• Interaction score based on PIE dataset

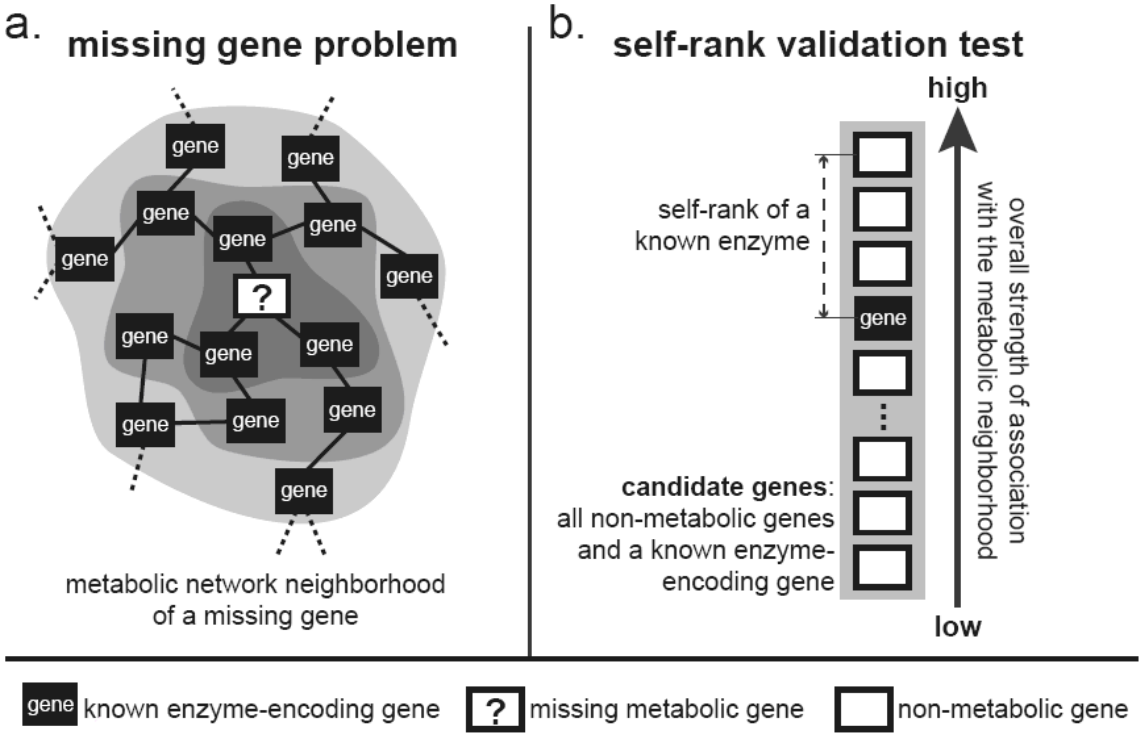


Figure C1.a. Illustration of the missing gene problem.
Figure C1.b. Illustration of the Self-rank validation test

Metabolic network neighborhood of a missing metabolic enzyme is shown. The neighborhood comprises layers with increasing radii (indicated by shading). Majority of the enzyme-encoding genes in the neighborhood are known. **b. Illustration of the self-rank validation test**. Ability to predict known enzyme-encoding genes is tested by measuring the rank of a true enzyme-encoding gene in the candidate set. The candidates are ordered according to overall strength of functional association with the metabolic network neighborhood of the enzyme. The set contains all genes that are not already part of the metabolic network.

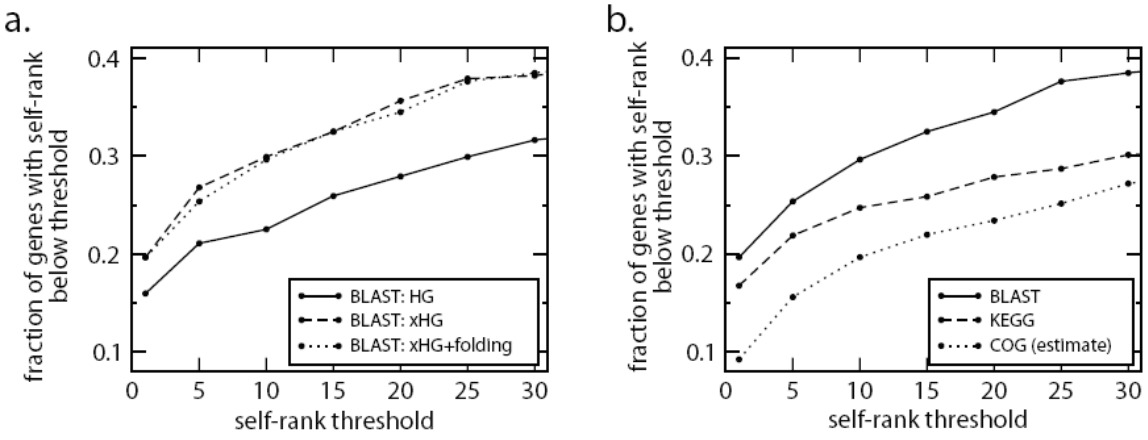


Figure C2a. Performance of different phylogenetic profile datasets and corrections..
Figure C2b. The self-rank performance of the 1st layer phylogenetic profile score

a. A cumulative self-rank distribution is shown for E. coli enzymes, as predicted based on the phylogenetic profile associations with the 1st layer of the metabolic network neighborhood. Performance of a regular hypergeometric distribution is shown (HG), together with extended hypergeometric (xHG) and folding (xHG+folding) corrections. The scores are calculated on the BLAST-based dataset. **b.** The self-rank performance of the 1st layer phylogenetic profile score, calculated using extended hypergeometric distribution with folding is shown for BLAST-based, KEGG-based and COG orthology datasets. The performance of the COG orthology dataset is corrected for the metabolic gene coverage bias.

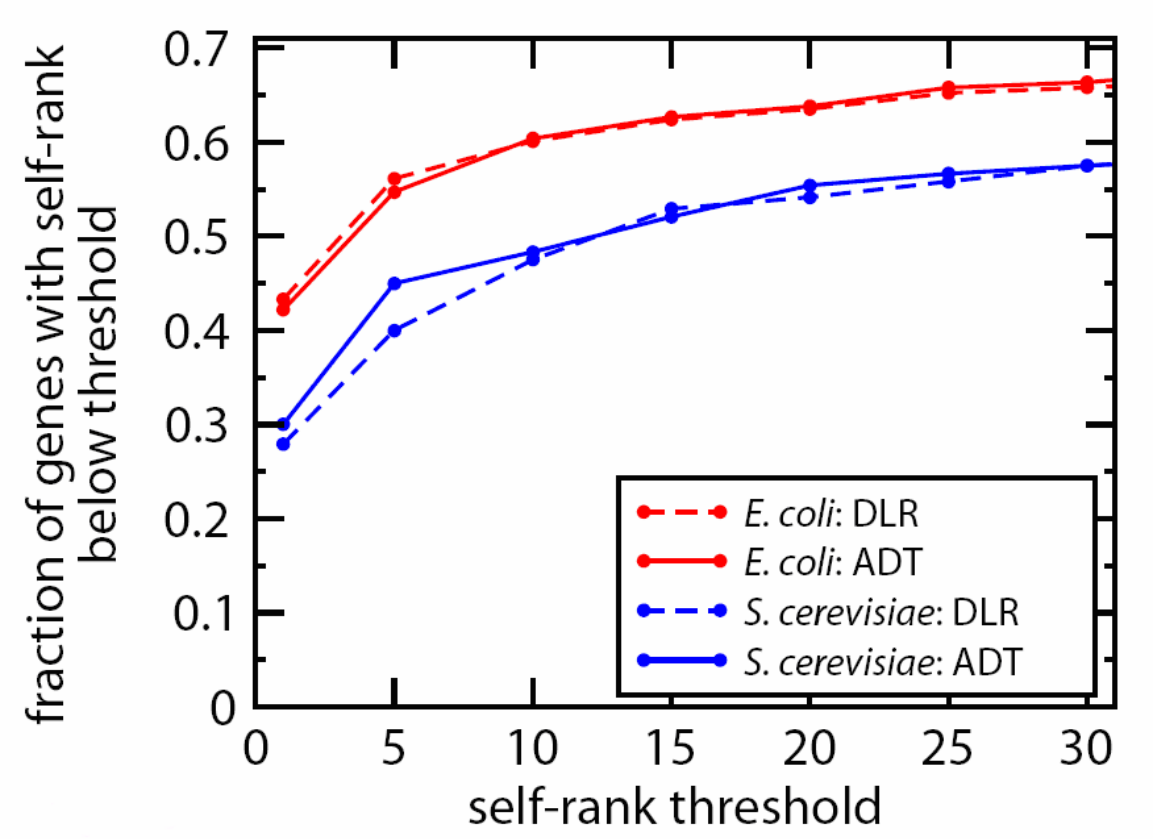


Figure C3. Comparison of ADT and DLR methods for combining multiple association evidence types.

Cumulative self-rank distribution is shown for *E. coli* and *S. cerevisiae* metabolic enzymes based on the combined association evidence (see Methods). The performance is compared for DLR (dashed curves) and ADT (solid curves) methods.

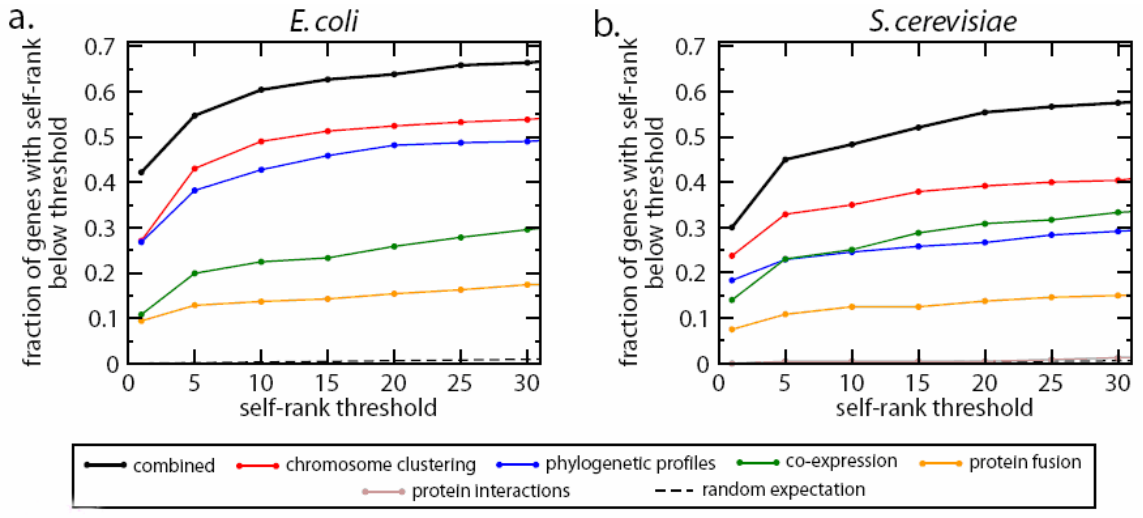


Figure C4. Enzyme predictions based on individual and combined types of association evidence.

Cumulative self-rank distribution is shown for the metabolic enzymes of **a.** *E. coli* metabolism and **b.** *S. cerevisiae* metabolism. Predictions are generated based on association with the first three layers of the metabolic network neighborhood, using ADT classifier with 10-fold validation.

6 Conclusion

The ideas in our 2001 DARPA grant proposal about design of complex synthetic biological polymers and software to aid this have coalesced into a vibrant field (e.g. the recent Dec 2005 issue of Nature has a collection of articles related to the new discipline of Synthetic Biology). Initial designs focused on biopolymer synthesis (DNA, RNA, protein), in vitro. That project was expanded with supplementary funding to include experimental work which was then transitioned out to a DOE GtL Center grant which resulted in a Nature paper and commercial licensing (to CodonDevices). The metabolic modeling (MOMA) has developed as very useful in stand-alone software and as an example of the successes and challenges on merging such software into the BioSPICE/BioCOMP community vision. The metabolic modeling has also been very successful and lives on in our Harvard/MIT DOE GTL Center. The BioSpice component led by Sri Kumar and others in our DARPA BioCOMP grant program aimed to develop interoperability and constituted a longer-term community-building exercise.

Technology Transition:

How the impact of this work is measured: Literature citations and milestones of licensees set by Harvard Medical School Office of Technical Licensing (HMS OTL) (Maryanne Fenerjian <maryanne_fenerjian@hms.harvard.edu>)

Prototype available for dissemination: In situ fluorescent base extension. Purpose: Identity and quantitation based on single DNA molecules. Environment requirements: Research laboratory. Point of contact / email address: Jay Shendure <jay_shendure@student.hms.harvard.edu> http://arep.med.harvard.edu/Polonator/

Systems available for dissemination:

MapQuant. Purpose: New Open-Source Software for Large-Scale Protein Quantitation Environment requirements: Research computers supporting Linux or Windows. Point of contact / email address: Kyriacos Leptos <leptos@fas.harvard.edu>
http://club.med.harvard.edu/MapQuant/

Minimization of Metabolic Adjustments (MOMA). Purpose: Optimization of metabolic network utilization in engineered (or mutant) genomes. Environment requirements: Research computers supporting Perl & C. Point of contact / email address: Daniel Segre <dsegre@genetics.med.harvard.edu>
http://arep.med.harvard.edu/moma/

7 Bibliography

Section 3

Andachi, Y., Yamao, F., Muto, A., and Osawa, S. (1989). Codon recognition patterns as deduced from sequences of the complete set of transfer RNA species in Mycoplasma capricolum. Resemblance to mitochondria. J Mol Biol 209, 37-54.

Bjork, G. R. (1995). Biosynthesis and function of modified nucleosides. In tRNA: structure, biosynthesis, and function (ASM Press, Washington D.C.).

Carr, P. A., Park, J. S., Lee, Y. J., Yu, T., Zhang, S., and Jacobson, J. M. (2004). Protein-mediated error correction for de novo DNA synthesis. Nucleic Acids Res 32, e162.

Cho, M. K., Magnus, D., Caplan, A. L., and McGee, D. (1999). Policy forum: genetics. Ethical considerations in synthesizing a minimal genome. Science 286, 2087-2090.

Culver, G. M., and Noller, H. F. (1999). Efficient reconstitution of functional Escherichia coli 30S ribosomal subunits from a complete set of recombinant small subunit ribosomal proteins. RNA 5, 832-843.

Dahl, F., Baner, J., Gullberg, M., Mendel-Hartvig, M., Landegren, U., and Nilsson, M. (2004). Circle-to-circle amplification for precise and sensitive DNA analysis. Proc Natl Acad Sci U S A 101, 4548-4553.

Del Campo, M., Kaya, Y., and Ofengand, J. (2001). Identification and site of action of the remaining four putative pseudouridine synthases in Escherichia coli. RNA 7, 1603-1615.

Diaconu, M., Kothe, U., Schlunzen, F., Fischer, N., Harms, J. M., Tonevitsky, A. G., Stark, H., Rodnina, M. V., and Wahl, M. C. (2005). Structural basis for the function of the ribosomal L7/12 stalk in factor binding and GTPase activation. Cell 121, 991-1004.

El Hage, A., Sbai, M., and Alix, J. H. (2001). The chaperonin GroEL and other heat-shock proteins, besides DnaK, participate in ribosome biogenesis in Escherichia coli. Mol Gen Genet 264, 796-808.

Forster, A. C., and Altman, S. (1990). External guide sequences for an RNA enzyme. Science 249, 783-786.

Forster, A. C., Cornish, V. W., and Blacklow, S. C. (2004). Pure translation display. Anal Biochem 333, 358-364.

Forster, A. C., and Symons, R. H. (1987). Self-cleavage of virusoid RNA is performed by the proposed 55-nucleotide active site. Cell 50, 9-16.

Forster, A. C., Tan, Z., Nalam, M. N. L., Lin, H., Qu, H., Cornish, V. W., and Blacklow, S. C. (2003). Programming peptidomimetic syntheses by translating genetic codes designed de novo. Proc Natl Acad Sci U S A 100, 6353-6357.

Forster, A. C., Weissbach, H., and Blacklow, S. C. (2001). A simplified reconstitution of mRNA-directed peptide synthesis: activity of the epsilon enhancer and an unnatural amino acid. Anal Biochem 297, 60-70.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., and et al. (1995). The minimal gene complement of Mycoplasma genitalium. Science 270, 397-403.

Fromont-Racine, M., Senger, B., Saveanu, C., and Fasiolo, F. (2003). Ribosome assembly in eukaryotes. Gene 313, 17-42.

Giege, R., Sissler, M., and Florentz, C. (1998). Universal rules and idiosyncratic features in tRNA identity. Nucleic Acids Res 26, 5017-5035.

Gitai, Z. (2005). The new bacterial cell biology: moving parts and subcellular architecture. Cell 120, 577-586.

Green, R., and Noller, H. F. (1996). In vitro complementation analysis localizes 23S rRNA posttranscriptional modifications that are required for Escherichia coli 50S ribosomal subunit assembly and function. RNA 2, 1011-1021.

Green, R., and Noller, H. F. (1999). Reconstitution of functional 50S ribosomes from in vitro transcripts of Bacillus stearothermophilus 23S rRNA. Biochemistry 38, 1772-1779.

Heurgue-Hamard, V., Champ, S., Engstrom, A., Ehrenberg, M., and Buckingham, R. H. (2002). The hemK gene in Escherichia coli encodes the N(5)-glutamine methyltransferase that modifies peptide release factors. EMBO J 21, 769-778.

Hutchison, C. A., Peterson, S. N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., Smith, H. O., and Venter, J. C. (1999). Global transposon mutagenesis and a minimal Mycoplasma genome. Science 286, 2165-2169.

Isaacs, F. J., Dwyer, D. J., Ding, C., Pervouchine, D. D., Cantor, C. R., and Collins, J. J. (2004). Engineered riboregulators enable post-transcriptional control of gene expression. Nat Biotechnol 22, 841-847.

Jaffe, J. D., Berg, H. C., and Church, G. M. (2004). Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 4, 59-77.

Kerner, M. J., Naylor, D. J., Ishihama, Y., Maier, T., Chang, H. C., Stines, A. P., Georgopoulos, C., Frishman, D., Hayer-Hartl, M., Mann, M., and Hartl, F. U. (2005). Proteome-wide analysis of chaperonin-dependent protein folding in Escherichia coli. Cell 122, 209-220.

Khaitovich, P., Tenson, T., Kloss, P., and Mankin, A. S. (1999). Reconstitution of functionally active Thermus aquaticus large ribosomal subunits with in vitro-transcribed rRNA. Biochemistry 38, 1780-1788.

Khan, S. A. (1997). Rolling-circle replication of bacterial plasmids. Microbiol Mol Biol Rev 61, 442-455.

Koonin, E. V. (2000). How many genes can make a cell: the minimal-gene-set concept. Annu Rev Genomics Hum Genet 1, 99-116.

Krzyzosiak, W., Denman, R., Nurse, K., Hellmann, W., Boublik, M., Gehrke, C. W., Agris, P. F., and Ofengand, J. (1987). In vitro synthesis of 16S ribosomal RNA containing single base changes and assembly into a functional 30S ribosome. Biochemistry 26, 2353-2364.

Kung, H.-F., Chu, F., Caldwell, P., Spears, C., Treadwell, B. V., Eskin, B., Brot, N., and Weissbach, H. (1978). The mRNA-directed synthesis of the alpha-peptide of beta-galactosidase, ribosomal proteins L12 and L10, and elongation factor Tu, using purified translational factors. Arch Biochem Biophys 187, 457-463.

Lewin, B. (2004). Genes VIII, 8th edn (Upper Saddle River, NJ: Pearson Prentice Hall).

Li, Z., and Deutscher, M. P. (1996). Maturation pathways for E. coli tRNA precursors: a random multienzyme process in vivo. Cell 86, 503-512.
Lietzke, R., and Nierhaus, K. H. (1988). Total reconstitution of 70S ribosomes from Escherichia coli. Methods Enzymol 164, 278-283.
Luisi, P. L. (2002). Toward the engineering of minimal living cells. Anat Rec 268, 208-214.
Maki, J. A., and Culver, G. M. (2005). Recent developments in factor-facilitated ribosome assembly. Methods 36, 313-320.
Mills, D. R., Peterson, R. L., and Spiegelman, S. (1967). An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. Proc Natl Acad Sci U S A 58, 217-224.
Mitra, R. D., and Church, G. M. (1999). In situ localized amplification and contact replication of many individual DNA molecules. Nucleic Acids Res 27, e34.
Mushegian, A. R., and Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci U S A 93, 10268-10273.
Nakahigashi, K., Kubo, N., Narita, S., Shimaoka, T., Goto, S., Oshima, T., Mori, H., Maeda, M., Wada, C., and Inokuchi, H. (2002). HemK, a class of protein methyl transferase with similarity to DNA methyl transferases, methylates polypeptide chain release factors, and hemK knockout induces defects in translational termination. Proc Natl Acad Sci U S A 99, 1473-1478.
Nierhaus, K. H., and Dohme, F. (1974). Total reconstitution of functionally active 50S ribosomal subunits from Escherichia coli. Proc Natl Acad Sci U S A 71, 4713-4717.
Nomura, M., and Erdmann, V. A. (1970). Reconstitution of 50S ribosomal subunits from dissociated molecular components. Nature 228, 744-748.
Ogle, J. M., and Ramakrishnan, V. (2005). Structural insights into translational fidelity. Annu Rev Biochem 74, 129-177.
Pohorille, A., and Deamer, D. (2002). Artificial cells: prospects for biotechnology. Trends Biotechnol 20, 123-128.
Richmond, K. E., Li, M. H., Rodesch, M. J., Patel, M., Lowe, A. M., Kim, C., Chu, L. L., Venkataramaian, N., Flickinger, S. F., Kaysen, J., et al. (2004). Amplification and assembly of chip-eluted DNA (AACED): a method for high-throughput gene synthesis. Nucleic Acids Res 32, 5011-5018. Print 2004.
Sauer, B. (2002). Cre/lox: one more step in the taming of the genome. Endocrine 19, 221-228.
Semrad, K., and Green, R. (2002). Osmolytes stimulate the reconstitution of functional 50S ribosomes from in vitro transcripts of Escherichia coli 23S rRNA. RNA 8, 401-411.
Semrad, K., Green, R., and Schroeder, R. (2004). RNA chaperone activity of large ribosomal subunit proteins from Escherichia coli. RNA 10, 1855-1860.
Service, R. F. (2005). How far can we push chemical self-assembly? Science 309, 95.
Shimizu, Y., Kanamori, T., and Ueda, T. (2005). Protein synthesis by pure translation systems. Methods 36, 299-304.
Soma, A., Ikeuchi, Y., Kanemasa, S., Kobayashi, K., Ogasawara, N., Ote, T., Kato, J., Watanabe, K., Sekine, Y., and Suzuki, T. (2003). An RNA-modifying enzyme that

governs both the codon and amino acid specificities of isoleucine tRNA. Mol Cell 12, 689-698.
Spirin, A. S., Baranov, V. I., Ryabova, L. A., Ovodov, S. Y., and Alakhov, Y. B. (1988). A continuous cell-free translation system capable of producing polypeptides in high yield. Science 242, 1162-1164.
Studier, F. W., Rosenberg, A. H., Dunn, J. J., and Dubendorff, J. W. (1990). Use of T7 RNA polymerase to direct expression of cloned genes. Methods Enzymol 185, 60-89.
Szostak, J. W., Bartel, D. P., and Luisi, P. L. (2001). Synthesizing life. Nature 409, 387-390.
Takyar, S., Hickerson, R. P., and Noller, H. F. (2005). mRNA helicase activity of the ribosome. Cell 120, 49-58.
Tian, J., Gong, H., Sheng, N., Zhou, X., Gulari, E., Gao, X., and Church, G. (2004). Accurate multiplex gene synthesis from programmable DNA microchips. Nature 432, 1050-1054.
Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T. S., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J. C., and Hutchison, C. A., 3rd. (1999). E-CELL: software environment for whole-cell simulation. Bioinformatics 15, 72-84.
Traub, P., and Nomura, M. (1968). Structure and function of E. coli ribosomes. V. Reconstitution of functionally active 30S ribosomal particles from RNA and proteins. Proc Natl Acad Sci U S A 59, 777-784.
Zhong, X. B., Lizardi, P. M., Huang, X. H., Bray-Ward, P. L., and Ward, D. C. (2001). Visualization of oligonucleotide probes and point mutations in interphase nuclei and DNA fibers using rolling circle DNA amplification. Proc Natl Acad Sci U S A 98, 3940-3945.
Zhou, X., Cai, S., Hong, A., You, Q., Yu, P., Sheng, N., Srivannavit, O., Muranjan, S., Rouillard, J. M., Xia, Y., et al. (2004). Microfluidic PicoArray synthesis of oligodeoxynucleotides and simultaneous assembling of multiple DNA sequences. Nucleic Acids Res 32, 5409-5417.
Zvereva, M. I., Shpanchenko, O. V., Dontsova, O. A., Nierhaus, K. H., and Bogdanov, A. 6 A. (1998). Effect of point mutations at position 89 of the E. coli 5S rRNA on the assembly and activity of the large ribosomal subunit. FEBS Lett 421, 249-251.

Section 4

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.
Asthana, S., O.D. King, F.D. Gibbons, and F.P. Roth. 2004. Predicting protein complex membership using probabilistic network reliability. *Genome Res* **14:** 1170-1175.
Bergmann, S., J. Ihmels, and N. Barkai. 2004. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol* **2:** E9.
Bishop, A.C., J. Xu, R.C. Johnson, P. Schimmel, and V. de Crecy-Lagard. 2002. Identification of the tRNA-dihydrouridine synthase family. *J Biol Chem* **277:** 25090-25095.

Bobik, T.A. and M.E. Rasche. 2001. Identification of the human methylmalonyl-CoA racemase gene based on the analysis of prokaryotic gene arrangements. Implications for decoding the human genome. *J Biol Chem* **276:** 37194-37198.

Borodina, I., P. Krabben, and J. Nielsen. 2005. Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism. *Genome Res* **15:** 820-829.

Bowers, P.M., M. Pellegrini, M.J. Thompson, J. Fierro, T.O. Yeates, and D. Eisenberg. 2004. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* **5:** R35.

Dudley, A.M., D.M. Janse, A. Tanay, R. Shamir, and G.M. Church. 2005. A global view of pleiotropy and phenotypically derived gene function in yeast. *Nature Molecular Systems Biology***:** doi: 10.1038/msb4100004.

Durand, D. and D. Sankoff. 2003. Tests for gene clustering. *J Comput Biol* **10:** 453-482.

Enright, A.J., I. Illopoulos, N.C. Kyrpides, and C.A. Ouzounis. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402:** 80-83.

Forster, J., I. Famili, P. Fu, B.O. Palsson, and J. Nielsen. 2003. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res.* **13:** 244-253.

Freund, Y. and L. Mason. 1999. The alternating decision tree learning algorithm. In *16th* I*nternational Conference on Mach*i*ne Learn*i*ng*, pp. 124-133.

Freund, Y. and R. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Computer and System Sci.* **55:** 119-139.

Gavin, A.C., M. Bosche, R. Krause, and P. Grandi. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415:** 141-147.

Green, M.L. and P.D. Karp. 2004. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bio*i*nformatics* **5:** 76.

Harkness, W.L. 1965. Properties of the extended hypergeometric distribution. *Annals of Mathematical Statistics* **36:** 938-945.

Ho, Y., A. Gruhler, A. Heilbut, G.D. Bader, L. Moore, S.L. Adams, A. Millar, and P. Taylor. 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415:** 180-183.

Hughes, T.R., M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, and H.A. Bennett. 2000. Functional discovery via a compendium of expression profiles. *Cell* **102:** 109-126.

Huynen, M., B. Snel, W. Lathe, 3rd, and P. Bork. 2000. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* **10:** 1204-1210.

Huynen, M.A. and P. Bork. 1998. Measuring genome evolution. *Proc Natl Acad Sci U S A* **95:** 5849-5856.

Huynen, M.A., B. Snel, C. von Mering, and P. Bork. 2003. Function prediction and protein networks. *Curr Op*in *Cell Biol* **15:** 191-198.

Ito, T., K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki. 2000. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A* **97:** 1143-1147.

Itoh, M., T. Akutsu, and M. Kanehisa. 2004. Clustering of database sequences for fast homology search using upper bounds on alignment score. *Genome* In*form Ser Workshop Genome* In*form* **15:** 93-104.

Jansen, R., H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, and M. Gerstein. 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302:** 449-453.

Kharchenko, P., G.M. Church, and D. Vitkup. 2005. Expression dynamics of a cellular metabolic. *Molecular Systems Biology*.

Kharchenko, P., D. Vitkup, and G.M. Church. 2004. Filling gaps in a metabolic network using expression information. *Bio*in*formatics* **20 Suppl 1:** I178-I185.

Lee, I., S.V. Date, A.T. Adai, and E.M. Marcotte. 2004. A probabilistic functional network of yeast genes. *Science* **306:** 1555-1558.

Lee, J.M. and E.L. Sonnhammer. 2003. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res* **13:** 875-882.

Maltsev, N., E.M. Glass, G. Ovchinnikova, and Z. Gu. 2005. Molecular Mechanisms Involved in Robustness of Yeast Central Metabolism against Null Mutations. *J Biochem (Tokyo)* **137:** 177-187.

Marcotte, E.M., M. Pellegrini, H.L. Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg. 1999. Detecting protein function and protein-protein interactions from genome sequences. *Science* **285:** 751-753.

Matic, S., S. Widell, H.E. Akerlund, and G. Johansson. 2001. Interaction between phosphofructokinase and aldolase from Saccharomyces cerevisiae studied by aqueous two-phase partitioning. *J Chromatogr B Biomed Sci Appl* **751:** 341-348.

Meredith, T.C. and R.W. Woodard. 2003. Escherichia coli YrbH is a D-arabinose 5-phosphate isomerase. *J Biol Chem* **278:** 32771-32777.

Middendorf, M., A. Kundaje, Y. Freund, C. Wiggins, and C. Leslie. 2005. Motif discovery through predictive modeling of gene regulation. *Proc. RECOMB***:** 538-552.

Middendorf, M., A. Kundaje, C. Wiggins, Y. Freund, and C. Leslie. 2004. Predicting genetic regulatory response using classification. *Bio*in*formatics* **20 Suppl 1:** I232-I240.

Osterman, A. and R. Overbeek. 2003. Missing genes in metabolic pathways: a comparative genomics approach. *Curr Op*in *Chem Biol* **7:** 238-251.

Overbeek, R., M. Fonstein, M. D'Souza, G.D. Pusch, and N. Maltsev. 1999. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96:** 2896-2901.

Pellegrini, M., E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96:** 4285-4288.

Press, W.H., S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. 2002. *Numerical Recipes* in *C++: The Art of Scientific Comput*in*g*. Cambridge University Press, Cambridge, UK.

Reed, J.L., T.D. Vo, C.H. Schilling, and B.O. Palsson. 2003. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biol* **4:** R54.

Schapire, R. 2002. The boosting approach to machine learning: An overview. *MSRI Workshop on Nonlinear Estimation and Classification*.

Schapire, R., Y. Freund, P. Barlett, and W.S. Lee. 1997. Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Stat.* **26:** 1651-1686.

Sherlock, G., T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J.C. Matese, S.S. Dwight, M. Kaloper, S. Weng, H. Jin, C.A. Ball, M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, and J.M. Cherry. 2001. The Stanford Microarray Database. *Nucleic Acids Res* **29:** 152-155.

Snel, B., V. van Noort, and M.A. Huynen. 2004. Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes. *Nucleic Acids Res* **32:** 4725-4731.

Tatusov, R.L., A.R. Mushegian, P. Bork, N.P. Brown, W.S. Hayes, M. Borodovsky, K.E. Rudd, and E.V. Koonin. 1996. Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with Escherichia coli. *Curr Biol* **6:** 279-291.

Tatusov, R.L., D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova, and E.V. Koonin. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29:** 22-28.

Teichmann, S.A. and M.M. Babu. 2002. Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends Biotechnol* **20:** 407-410; discussion 410.

Troyanskaya, O.G., K. Dolinski, A.B. Owen, R.B. Altman, and D. Botstein. 2003. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc Natl Acad Sci U S A* **100:** 8348-8353.

Uetz, P., L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J.M. Rothberg. 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403:** 623-627.

van Noort, V., B. Snel, and M.A. Huynen. 2003. Predicting gene function by conserved co-expression. *Trends Genet* **19:** 238-242.

von Mering, C., M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel. 2003. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* **31:** 258-261.

Wong, S.L., L.V. Zhang, A.H. Tong, Z. Li, D.S. Goldberg, O.D. King, G. Lesage, M. Vidal, B. Andrews, H. Bussey, C. Boone, and F.P. Roth. 2004. Combining biological networks to predict genetic interactions. *Proc Natl Acad Sci U S A* **101:** 15682-15687.

Wu, J., S. Kasif, and C. DeLisi. 2003. Identification of functional links between genes using phylogenetic profiles. *Bio*in*formatics* **19:** 1524-1530.

Yamanishi, Y., J.P. Vert, and M. Kanehisa. 2004. Protein network inference from multiple genomic data: a supervised approach. *Bio*in*formatics* **20 Suppl 1:** I363-I370.

Yamanishi, Y., J.P. Vert, and M. Kanehisa. 2005. Supervised enzyme network inference from the integration of genomic data and chemical information. *Bio*in*formatics* **21 Suppl 1:** i468-i477.

Yanai, I., A. Derti, and C. DeLisi. 2001. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc. Natl. Acad Sci.* **98:** 7940-7945.

Yanai, I., J.C. Mellor, and C. DeLisi. 2002. Identifying functional links between genes using conserved chromosomal proximity. *Trends Genet* **18:** 176-179.

8 Publication List

Aach JA and Church GM. (2004) Mathematical models of diffusion-constrained polymerase chain reactions: basis of high-throughput nucleic acid assays and simple self-organizing systems. J Theor. Biol. May 7;228(1):31-46.

Forster, AC & Church, GM (2006) Synthesizing a Minimal Cell (submitted)

Jaffe, JD, Stange-Thomann, N, Smith, C, DeCaprio, D, Sheila Fisher, S, Butler, J, Calvo, S, Elkins, T, FitzGerald, MG, Hafez, N, Kodira CD, Major J, Wang S, Wilkinson, J, Nicol, R, Nusbaum, C, Birren, B, Berg, HC, Church GM (2004) The complete genome and proteome of Mycoplasma mobile. Genome Res. 2004 Aug;14(8):1447-61.

Kharchenko, P, Church, GM, Vitkup, D. (2004) Filling gaps in a metabolic network using expression information. Bioinformatics. 2004 Aug 4;20 Suppl 1:I178-I185.

Kharchenko, P., G.M. Church, and D. Vitkup. 2005. Expression dynamics of a cellular metabolic. Molecular Systems Biology. doi:10.1038/msb4100023

Lelandais, G, Le Crom,S, Devaux, F, Vialette, S, Church, GM, Jacq, C and Marc,P (2003) yMGV : a cross-species expression data mining tool. Nucleic Acids Res. 90001 D323-D325.

Leptos, KC, Sarracino, DA, Church, GM (2006) MapQuant: A New Open-Source Software for Large-Scale Protein Quantitation Proteomics. 2006 Mar;6(6):1770-82.

Mikkilineni V, Mitra RD, Merritt J, DiTonno JR, Church GM, Ogunnaike B, Edwards JS. Digital quantitative measurements of gene expression. Biotechnol Bioeng. 2004 Apr 20;86(2):117-24.

Segre, D, Zucker, J, Katz, J, Lin, X, D'haeseleer, P, Rindone, W, Karchenko, P, Nguyen, D, Wright, M, and Church, GM (2003) From annotated genomes to metabolic flux models and kinetic parameter fitting. Omics 7:301-16.

Tian J, Gong H, Sheng N, Zhou X, Gulari E, Gao X, & Church GM (2004) Accurate Multiplex Gene Synthesis from Programmable DNA Chips. Nature. 2004 Dec 23;432(7020):1050-4.