Minireview

Genomics & Informatics Vol. 2(1) 7-18, March 2004

Computational Challenges for Integrative Genomics

Junhyong Kim[1,2]* and Paul Magwene[1]

[1]Department of Biology,
[2]Department of Computer and Information Science,
Penn Center for Bioinformatics,
University of Pennsylvania,
415 S University Ave, Philadelphia, PA 19104-6018, USA

Abstract

Integrated genomics refers to the use of large-scale, systematically collected data from various sources to address biological and biomedical problems. A critical ingredient to a successful research program in integrated genomics is the establishment of an effective computational infrastructure. In this review, we suggest that the computational infrastructure challenges include developing tools for heterogeneous data organization and access, innovating techniques for combining the results of different analyses, and establishing a theoretical framework for integrating biological and quantitative models. For each of the three areas - data integration, analyses integration, and model integration - we review some of the current progress and suggest new topics of research. We argue that the primary computational challenges lie in developing sound theoretical foundations for understanding the genome rather than simply the development of algorithms and programs.

Keywords: Integrative genomics, computational biology, bioinformatics, probabilistic modeling

Introduction

The era of genomics was brought into prominence with the initiation of the Human Genome Project in October of 1990. The Human Genome Project was declared complete in April 2003 after 13 years of international effort (Collins, Morgan, and Patrinos, 2003). Many other genome projects have been completed with more than 100 (non-viral) whole genomes already available and many more genomes in the pipeline. Current sequencing capacity is such that a relatively medium-sized genome like that of *Drosophila pseudoobscura* can be completed by a single company, essentially in one month. As genomic sequencing has become routine, attention has begun to focus on so-called post-genomic technologies including large-scale transcript assays, proteomics, combinatorial chemistry, and so on (with an unfortunate accompanying proliferation of neologisms including: transcriptome, proteome, metabolome, etc.). It is too premature to put strong emphasis on any particular sub-field such as "proteomics" or "functional genomics" but it is clear all of these activities involve an important principle: large-scale, high-throughput, systematic collection of data without explicit predefined hypotheses. In this review, we will refer to all such activities "genomic sciences" or "genomic approaches."

As mentioned, genomic approaches to biological problems involve the idea of systematic data collection at all levels of information, including the genome, the state of the cell (e.g., the transcriptome, the proteome, the metabolome, etc.), and the phenotype (e.g., anatomy, physiology, clinical, etc.). Genomic approaches have several important advantages. First, they leverage high-throughput technologies and streamlined production processes to obtain data at a far higher speed and lower expense than is possible through individual experimentation. Second, by generating a large collection of data and making it publicly available, they allow the research community access to resources that are relevant to the original problem as well as to many innovative and unanticipated problems. Most importantly, the availability of large-scale systematic data allows for new kinds of inferences and discoveries that can be only made when the entire data is available. For example, in recent years, genomic technologies have been increasingly used to address human disease specific problems such as typing hidden variation in human cancer (e.g., Alizadeh *et al.*, 2000; West *et al.*, 2001), tracking infectious disease agents and hosts (e.g., Metzker *et al.*, 2002; Hillis, 2000), discovering new drug targets (e.g., Foth *et al.*, 2003; Kissinger *et al.*, 2002), discovering new genes and molecular processes involved in diseases, and many other innovative applications.

The profusion of various large-scale data has led to the coining of the phrase "Integrated Genomics." This phrase, as used in various literatures, is associated with a vague (but exciting) hope that, given large quantities

* Corresponding author: E-mail Junhyong@sas.upenn.edu, Tel +1-215-746-5187, Fax +1-215-898-8780
Accepted 15 March 2004

of genome-scale data collected at different levels of biological organization and from a variety of organisms, we should be able to utilize all this information in a comprehensive manner.

The grand answer to such a challenge "here is a way that anybody asking any biological question can bring all evidence to bear" is obviously impossible and really speaks to how science is carried out in general. However, from a computational viewpoint, we might envision an integrated computing environment for genome-scale modeling in which a scientist would be able to describe a hypothesis to be tested as a precise combination of model components. Each specification would be ***computable*** and would be coupled with appropriate data sources to score the hypothesis. The computable components would be modular such that different ideas or data expressed as these modules could be interchangeably combined to construct a grand computable hypothesis and test of hypothesis. For example, one should be able to specify various hypotheses about signal transduction pathways, which in turn could be used to compute a genome-wide expression pattern, this expression pattern serving as input into a model relating the expression pattern to phenotypic data.

This, of course, is a rather lofty goal given that so many of the biological and computational questions relevant to building such a system are still unanswered. Rather than focusing on the details of what such a system might look like, we instead focus our discussion on a number of "integrative principles" that might be employed as we work towards such a goal. At the most basic level we need tools and systems for ***Data Integration*** including signal processing of the primary data (e.g., microarray measurements), statistical characterization (e.g., image analysis from radiological measurements, text mining), and organization (e.g., integrated databases). At the next level we need ***Analysis Integration*** wherein we can employ a multitude of analysis tools for biological inference but in such a way that different analyses and/or their outputs can be combined. Finally, we need tools and systems for ***Model Integration***. Ideally, the various analysis tools should incorporate biological process models as their principles. Similarly, their outputs should either explicitly or implicitly suggest new biological process models (e.g., an estimate of gene regulatory network should then translate into a new model of molecular interactions and pathways). These models then should be combinable either in their output states or as principles for analyses. For example, two different models for a gene family, say one based on a Hidden Markov Model and another based on secondary structure, should be combinable for function prediction.

The goal of this review is to explore some of the computational challenges within each of these three areas that must be met if we are to practice a truly integrative genomics. In particular, we discuss our vision of the principles and accompanying computational framework that would allow for a theoretically sound, uniform approach to integrated genomic analyses. In addition to discussing these integrative principles, we highlight current research and future challenges for establishing a computational infrastructure to meet the demands of integrated large-scale data analysis.

Data Integration

The ultimate goal of data integration in bioinformatics is to provide analysis tools with unified access to the great variety of large-scale data. As mentioned in the introduction, there are three classes of problems that need to be solved by a computational system for data integration. The first class of problems is what we might call "instrumentation problems." These are computational problems associated with obtaining the most accurate and efficient measurements from organisms. The second class of problems might be called "data characterization problems." These problems involve transforming and encapsulating primary data in an organized and comparable manner. For example, taking functional MRI data and identifying morphological landmarks in such a way to make different individual measurements comparable is a data characterization problem (e.g., Kennedy *et al.*, 2002). The third class of problems is the more familiar problem of establishing a database system for easy access to heterogeneous data.

Data Acquisition

The main tradeoff in many high-throughput technologies is that between volume of data and accuracy of data. Some high-throughput technologies such as standard multiplexed sequencing have been sufficiently standardized such that there is no real tradeoff between data volume and data quality. However, for the majority of newer technologies such as transcript assays, proteomics, microarray biochemistry, and sequencing by hybridization, obtaining reliable instrumentation is still a large challenge. An example of this is the computations required to obtain reliable microarray gene expression measurements. For cDNA type of arrays, many different methods have been presented for correcting for array-to-array standardization, dye/channel bias, robotic pin bias, and so on (e.g., Kerr and Churchill, 2001; Yang *et al.*, 2002). In practice, all of these factors play a great role in the array measurements; yet there are still significant sources of error and there is a need to develop new computational tools especially with respect to noise reduction and feature recognition at the primary image level (e.g., de-trend analysis). Similar problems arise for 2D protein gels as well as for primary signal processing from a mass spectroscopy instrument (Efrat *et al.*, 2002). As a third example, detection of genomic alteration by array Comparative Genomic Hybridization (aCGH) is quickly becoming an essential tool for cancer biology. However, there are many computational and statistical challenges associated with aCGH, such as reliable copy number determination and extrapolation of the clone data to genomic regions (Mantripragada *et al.*, 2004; Wang and Guo, 2004).

It is expected that many of these problems will be solved soon with a combination of better physical instruments and algorithms. For example, the early automated sequencers also had many instrumentation problems with both the device itself and the algorithms that interpreted the spectra. However, there is also another important family of problems that require more research, namely computational support for developing new high-throughput technologies. For example, Sequencing By Hybridization (SBH) is an important strategy for reaching the next level of efficiency in genomics (Ben-Dor *et al.*, 2001). The design of a proper set of oligomers that will minimize cross hybridizations and allow efficient assembly is still an unresolved problem. In particular, many of the current approaches do not seem to take into account that much of the genome consists of related gene families or related sequence pieces. This kind of computational support for data acquisition is a weakly developed area that is likely to be critical for continued development of genomic technologies. Another case example can be found in computational support for phenotyping. Ultimately, the real interest in genomic approaches is in associating the rapidly obtained genomics data to phenotypic predictions such as risk for cancer recurrence. High-throughput phenotyping, especially at the anatomical level, is a huge bottleneck in allowing sufficient data collection. Image analysis, image reconstruction, and image-based signal processing are all expected to be important areas of computational development (Yarrow *et al.*, 2003) in this context.

Data Characterization and Encapsulation

In information processing theory, a distinction is made between "data" and "information", where data refers to the primary measurements (after correction for instrumentation problems) and information refers to a transformation of the data into an organized structure that can be used for higher level inferences (so-called "knowledge"). A canonical problem in transforming data to information is processing of text found in scientific literature. From the computational point of view, the text is unstructured data that must be organized into specifically typed information. For example, we may be interested in collecting information to construct an overall view of *Drosophila* development. Each primary text must be mined for gene names, mutations, possible phenotypic effects, possible gene interactions, and so on. The steps involved are: (1) developing annotation standards and software tools for annotating training documents with respect their syntactic structure and semantic content (especially tied to specific domain knowledge, say *Drosophila* development); (2) curating a training set following the syntactic rules; and (3) developing algorithms for automatic training of components that will recognize the relevant syntactic and semantic structures accurately in new documents. In many ways, these problems point to a model driven (or more generically, knowledge driven) characterization of the primary data. That is, to determine the important components of a journal article, we already need some model of the knowledge ontology. We can analogize this to similar ideas in mathematical statistics where the idea of a "sufficient statistic" is to obtain a function of the data that is sufficient to characterize some probability distribution (i.e., a model). In other words, data characterization involves transformation and reduction of data to the most relevant bits of biological information. But, we cannot determine what is most relevant without a prior knowledge model.

Using programming jargon we might call this process of transforming and extracting primary data to a representation that provides sufficient and efficient information vis-a-vis some prior model (or even a vaguely determined set of prior models), ***data encapsulation***. We want our data encapsulated in such a way that only the interface necessary for knowledge modeling in any particular domain is exposed. An important aspect of encapsulation is what we might call "context independence". Given some primary measurements, we would like to extract information that is as context independent as possible. For example, one might consider expression microarray normalization as a process of making the measurement values context independent of particular arrays. The act of attaching functional annotation to a particular subsequence of the genome, say calling it an exon, is also an act that makes a particular stretch of sequence acquire a context independent property, namely that of an exon.

A dual notion to making data context independent is making data "comparable". By this we mean characterizing the data in such a way that the representation in one measurement is comparable to another measurement. A concrete example of this is sequence alignment of multiple genomic strings. Sequence alignment makes the positions comparable (positional homology) such that we can compute quantities like distances between the two strings. The alignments also provide context independence for individual nucleotides in the sense that we can commonly discuss particular nucleotides independent of their context within each string. We return to a discussion of context independence and comparability in the section on Model Integration.

As mentioned previously, phenotyping is an important component of integrative genomics. A relatively straightforward phenotyping problem is the characterization of anatomical image data. Here we first have the challenge of encapsulating a large amount of information, i.e., pixels, in a biologically relevant manner. Suppose the data of interest are images of the brain. Our prior knowledge model includes the general geometry of the brain (oval 3D structure) and information about gross anatomy. This knowledge must be used to segment the image (encapsulate the data) into relevant anatomical regions. Given such segmentation we would next want to allow comparison between different brain images (sampled from different individuals or at different times), which might require isolating homologous landmarks, coordinatizing the landmarks, and transforming different coordinates from different images into a common system (Bookstein, 1991). Many phenotypic data characterization problems are even more complex. For example, in neurogenomics we are interested in genomic basis of complex phenotypes some of which may be only measured through behavioral assays such as memory, navigation, conditioning, and so on. In these cases, the very basic representation of data is a challenge. Many of these problems are similar to challenges that are found in systematics and taxonomy. Taking advantage of the many methods and ideas that have been developed in these literatures will aid greatly in our task of moving towards an integrative genomics.

Data Management and Accessibility

The last class of problems in data integration is the more common problem of establishing databases and connecting heterogeneous data sources. While there has been great progress in databases for genomics, this area continues to remain a major challenge (Stein, 2003). Much of the information associated with the Human Genome Project (HGP) and similar projects on other organisms is served through controlled Web interfaces rather than by direct access to conventional databases. Many difficulties arise from a lack of coordinated vocabularies. Recent efforts on developing ontologies will certainly help with this problem (Ashburner *et al.*, 2000; Harris *et al.*, 2004). A serious challenge is to enable biologists to cope with frequent changes of assumptions, experimental hypotheses, and techniques; biological data sources are often loosely coupled with voluntary quality and standards control and not uncommon introduction of completely novel data types. The traditional data integration scenario starts with the *local* schemas of several actual data sources and one *global* schema against which queries are asked. In the *warehousing* approach a database corresponding to the global schema is built. Warehousing, and more generally building derived value-added databases, is an important aspect of the landscape of biological data. For example, the GUS (Genomics Unified Schema) developed at the University of Pennsylvania is a warehousing scheme designed to integrate biological sequences, annotations, gene expression, gene regulation, and proteomics under the central dogma of biology: Genes to RNA to Protein (Davidson *et al.*, 2001). An important principle here is that by modeling the data scheme around the most invariant central theory for the types of data, we are able to derive consistent schema for heterogeneous data. By itself warehousing is not a satisfactory solution in the integration scenarios we consider here because there is too much volatility and need for flexibility (Stein, 2003). What is required is a set of approaches that promote *peer data integration*, where there is no global schema but a mapping between various different data schema. An especially important challenge for peer data integration is tools for dealing with data redundancy (Deutsch and Tannen, 2003) and semi-automatically discovering data relations by schema matching and mapping (Rahm and Bernstein, 2001).

Analysis Integration

The motivating observation behind analysis integration is that biological phenomena are commonly context dependent and any single inference based on marginal measurements tends to be weak with both large false positives and false negatives. For example, in Kim *et al.* (2000) we developed an algorithm for predicting novel multi-transmembrane proteins. However, application of this algorithm to genomic data can result in 10% of the potential genes scoring positive, which would translate into thousands of genes for the Drosophila genome. Overcoming problems like this requires incorporating additional types of analyses such as clustering candidates into gene families, considering codon usage, negative selection against known databases, and so on. Ideally, integration of all of these separate procedures should be as seamless as possible.

In order to achieve such seamless integration between different methods our analytical tools need to meet at least three criteria. First, the output of different analyses must be compatible in the sense that they must allow conjunction of the results. Where existing software tools do not provide for easy conjunctions, translators must be designed to make the output of various algorithms compatible. For example, the output of two different programs for exon prediction could be either sub-sequences or categorical labels (e.g., yes/no) over the (partially) shared input set. Thus the outputs can be combined because the programs share the same type of output or because they share the same type of inputs and the outputs are functions over the input types. Second, there must be a value scheme for quantifying the reliability of each output, such as probabilities, ranks, scores, and so on. Finally, there must be a way of computing a reasonable function (say, some calculus) of combined value schemes in order to characterize the acceptability of given hypotheses under combined analyses. The last two components allow us to combine heterogeneous algorithm/analyses results. The most natural approach is to adopt a probabilistic interpretation of the value scheme, which will allows us to use existing machinery from probability theory for the value calculus.

We note that combining inference from different analyses requires the consideration of interactions of different biologically-based analysis schemes, not just some algorithmic scheme for combining analyses. For example, given a difficult optimization problem such as the maximum likelihood phylogeny estimation (cf, Felsenstein, 2003) one might create a strategy of combining the results of various heuristic algorithms. However, this is different from the problem of maximizing the probability of obtaining the correct phylogeny, which might be best approached by analyzing different data sources as well as incorporating any constraints from prior biological knowledge. Thus, any value scheme attached to the results should be based on some reasonable principle of relating the output to the biological goal, and not just on algorithmic considerations.

Designing value schemes

Two of the most important ways in which values could be assigned to outputs are the following:

1) A stochastic model is chosen for the process producing the input data. The value associated with an output is then the probability that the particular output is produced. For example, in a pairwise alignment problem, we could define an underlying stochastic process that transforms an unknown common string into the two given strings (or, equivalently, one of the given strings into the other) with specific rates of insertions, deletions and substitutions. The value or score associated with an alignment would then be the probability that the alignment represents a homology of the process. Actual alignment scoring schemes in use do correspond, modulo some cosmetic translation, to this process view. As another example, in phylogeny estimation a stochastic model such as the Jukes-Cantor model or the Kimura 2-parameter model is chosen (Felsenstein, 2003). The likelihood of a particular tree topology with particular edge lengths is proportional to the probability that this choice would produce the observed data (assuming that the prior probabilities are uniform).
2) The value associated with an output is its significance, i.e., the (un)likelihood of observing this output pattern in a suitable null model. The null model for DNA sequences could be sequences generated as *i.i.d.* random variables, sequences sampled uniformly from a language generated by a suitably defined grammar, sequences produced by simulating a suitable evolutionary process (if the evolutionary relationship between the taxa is understood), etc. For other objects such as RNA or protein secondary structures appropriate null models have to be defined based on the context.

We review some examples of model-based and null-hypothesis based value schemes below.

Model-based value schemes

Bioinformatics algorithms need to embody more explicit generative models of the data if their results are to be integrated following sound probabilistic principles. One example of where this has already been done is the use of hidden Markov models (HMMs) for biological sequence modeling and alignment (Durbin *et al.*, 1998). Heuristic alignment programs like BLAST can be seen as approximations of the more expensive full dynamic programming algorithms used with HMMs. However, HMMs in their basic form do not constitute a sufficiently flexible structure for analysis combination. Fortunately, HMMs are members of the broader class of probabilistic graphical models (Lauritzen, 1996) that includes Bayes' nets, Markov random fields, and (under appropriate interpretation) probabilistic grammars. One problem is that typical models with sufficient flexibility can be complex with large numbers of parameters. These parameters can be difficult to estimate and even if the parameter values are known the model can be difficult to compute. Thus, one key challenge is how to efficiently compute the probabilities. There are several possible approaches, including approximations, compacting (computing equivalence classes), computing marginal distributions, and using conditional distributions.

The problem of maximum likelihood estimation of phylogenies has been considered a notoriously difficult computational problem in the bioinformatics literature (reviewed in Felsenstein, 2003). Part of the argument is that even for a fixed choice of parameter values, computing the probability of a particular set of observations at the tips requires exponential time if done in a brute-force manner. By setting up an appropriate metric space for the set of possible models, analyzing the structural properties of the probability landscape over this metric space, and using the well-developed theory of approximation algorithms, one can show that it is possible to produce guaranteed approximations to the maximum likelihood for simple stochastic models of evolution (Farach and Kannan, 1999; Cryan *et al.*, 1998). One would like to extend these ideas to produce all choices of parameter values that produce high probabilities for the observed outputs and extend them to more complex model families.

Several general schemes are in common use for obtaining marginal probabilities from stochastic processes relatively efficiently including Markov Chain Monte Carlo (MCMC), Gibbs sampling, and Expectation Maximization (EM). However, actual efficiency and implementation is case-dependent. Conditional distributions are also commonly used in computing random variables (Gentle, 2003). The idea here is to find an easily computable distribution that is a dominating distribution for the domain of the random variable we desire. The desired distribution may be computed conditional on the dominating distribution. Application of all of these techniques to computational biology problems is still in infancy.

Null hypothesis-based value schemes

BLAST is perhaps the most famous program that produces value schemes in its output that rely on a null hypothesis. The null hypothesis used by BLAST is a fairly simple one – sequences generated position by position in *i.i.d.* fashion. In some applications an overly simple null hypothesis can artificially lower the "background signal" level thereby causing us to misread too many things as strong signals, and thus become inundated with too much signal. As an example, the BLAST statistics do not reflect the fact that all of the sequences in the databases have a dependency structure imposed by their evolutionary history. Thus if the researcher would like to assess the probability of getting a hit of score X if one were to search all molecules in life (as opposed to all random sequences), the resulting p-value would be quite different. The algorithmic difficulty with a more sophisticated null hypothesis is that computing the probability can be very hard. For instance, one can easily generate the null distribution that corresponds to a Jukes-Cantor process, but uniformly sampling strings of a certain length generated by a grammar is a difficult problem.

Using value schemes to integrate algorithmic analysis

As mentioned above, once analysis algorithms produce probability scores for the results vis-a-vis the focused biological inference, standard probability calculus can be used to derive a combined inference. The key is constructing a reasonable manner by which joint probabilities can be considered. For example, given two pairs of genes A and B, we may wish to estimate whether they have direct molecular interactions. Analysis of a functional genomics data set with protein-protein (P-P) interaction information may yield some estimate of Prob(A and B interact| positive P-P assay scores). Another analysis of their transcriptional co-expression may yield another probability score P(A and B interact | expression correlation 〉 0.5). The question is how to compute Prob(A and B interact | positive P-P scores AND expression correlation 〉 0.5). This requires us to model certain joint distributions and in a more complicated set of analyses, postulate a dependency relationship as the full joint distribution of all outcomes. A possible example of such combined inference is so-called Bayesian integration methods (Drawid and Gerstein, 2000; Troyanskaya *et al.*, 2003). This method involves selecting a prior assumption on marginal distributions (i.e., the prior distributions) and an inter-relational structure possibly represented as a graphical model. One problem with this approach is that in many cases the prior structure for the relationship of the joint events is extremely poorly known. It is difficult to introduce reasonable assumptions about the inter-relationships of many outcomes. For example, it is very difficult to state with confidence the joint probability of positive outcomes for two different interaction assays for two genes, say from proteomics and from expression analysis.

A possible approach here is to estimate the dependent relationships through a training set. This would be analogous to typical network modeling of gene regulation. For example, we may wish to predict whether a given open reading frame codes for G protein-coupled receptors (GPCR; a particular kind of protein). We may have a training data set with known GPCRs, several different prediction algorithms (e.g., Kim *et al.*, 2000) with probability scores. For each input sequence, each prediction algorithm produces a marginal probability value of GPCR assignment. Again, the relationship between algorithm outputs and the combined output can be represented as a probabilistic graphical model. We can then estimate the graph structure and conditional probabilities that maximize the likelihood of training dataset assignment at a special "combined output" vertex (Friedman and Koller, 2003). Possible problems here mirror those of standard network models, that is, computational complexity is high and it is difficult to obtain sufficient amounts of data.

Integrating different parametric models through a connecting semi-parametric model criteria

Process models are often integrated into analysis algorithms in the construction of the objective function or distance measures. For example, a model of molecular evolution might be incorporated into a likelihood objective function or distance measures between two sequences. Each objective function or distance measure with a finite set of parameters might be seen as representing a parametric process model. One approach to integrating different models is by connecting the different objective functions or distance measures into a semi-parametric family of models. Suppose we have k different objective functions or distance measures, $f_1(D, p_1),..., f_k(D, p_k)$, where D is the input data and P_i is the ith parameter set. A simple semi-parametric function is a weighted linear combination $L(D,\alpha\ , p) = \alpha_1 f_1(D, p_1)+\cdots+\alpha_k f_k(D, p_k)$, where α_1 is the ith mixture coefficient. This construction has the desirable property that we can recover the ith marginal objective function by setting all $\alpha_k = 0$, $k \neq i$.

An example of combining models through semi-parametric criteria comes from recent work in phylogenetic methods. J. Kim and M. Sanderson (unpublished) constructed such a semi-parametric model for phylogeny reconstruction where the semi-parametric function spanned the different models from the standard maximum parsimony estimator to the standard maximum likelihood estimator. In this case, the maximum parsimony estimator had an objective function of the form $f_{mp}(D, p_1 \cdots p_n)$ while the maximum likelihood had the form $f_m(D,\ p_1 = p_2 =\ \cdots p_n)$. That is, the maximum likelihood estimator had an identical objective function but maximized over a subspace of possible parameter set. This led to a semi parametric estimator of the form $L(D, \alpha\ , p) = f_{mp}(D, p_1\ \cdots\ p_n)+\alpha \cdot Var(p_1 \cdots p_n)$, thus when $\alpha\ = 0$, it led to the standard maximum parsimony model whereas when $\alpha = \infty$, it forced the variance term to zero resulting in the standard maximum likelihood estimator. This is a special case of different models where the models differ by subspace restrictions on the parameter set.

The key to this kind of model integration by semi-parametric objective function construction is selecting the appropriate connecting function. While a linear connecting function is simple, there are several desirable properties to consider and remain as open problems:

(1) Given some general form for the connecting function, $L(D, \alpha, p)$, and a reasonable definition of function complexity, say number of parameters, what is the least complex function such that there are values of α that recovers each marginal objective function? This problem asks us to construct a connecting function in a minimal way such that we can set the parameters of the connecting function to recover the properties of each input objective function. Thus, we are guaranteed that the semi-parametric model "contains" all the input models.

(2) For objective functions, we are typically interested in the characteristics of the points of the parameter set (or functions of the points) at the extrema. For example, if the objective function is the likelihood function for a phylogeny given some model of molecular evolution, then we are interested in what kinds of trees are output for some input dataset; that is, what kinds of trees maximize the likelihood function. If we were to construct a semi-parametric objective function, then we would be interested in the trees at the maximum of the new objective function. Suppose now we were given some collection of objective functions (and implied models) and a set of data. Also, suppose we had a desired output for each input data, say as a training set. Is there an efficient way to construct a semi-parametric connecting objective function such that the function has parameter values at which the training set can be recovered? That is, if we are given some choice of models and a set of desired outputs for an input data set, can we construct a family of integrated models that will recover the outputs? This question asks whether a method can be found to constructively generate a semi-parametric method that might agree with some prior biological knowledge.

(3) The semi-parametric connection function is a family of models in which depending on some parameter set, which we denoted as α , we can change the behavior of the analysis making it more like one model

versus another. In biological situations, this change in behavior may be scale or process dependent. For example, when comparing the expression of two genes, at low levels of expression we might consider their behavior similar because both are nearly "off". However, at higher levels of expression one might say the genes are related if the expression levels are correlated over different experiments. One simple semi-parametric implementation of this scenario is as $D(x, y) = \alpha\ |x - y| + (1 - \alpha)\cos^{-1}(x, y)$. That is, as a weighted combination of the Euclidean distance between the two expression levels x and y and the correlation between x and y (the angle). Biological reasoning suggests that the mixture parameter α should be related to the input values themselves. This suggests a new class of semi-parametric objective model families where the choice of the particular function is selected by the input data themselves. Similar problems might be raised with the training data set idea discussed above. Can we construct a semi-parametric model that chooses the appropriate member of the family from the characteristics of the input data? Studying this class of semi-parametric models is likely to lead to other problems such as whether one might be able to construct a family that can be consistently trained.

Model Integration

In a biological context the term 'modeling' generally refers to an abstraction of a biological system amenable to analysis for understanding and eventual prediction of the systems behavior via analytical or numerical computation. Many of the best computational analysis tools implicitly and explicitly incorporate models of biological processes including biophysical, biochemical, physiological, morphological, and evolutionary processes. For example, stochastic models of molecular evolution are used in likelihood-based estimation of phylogenies and known splicing reactions are used in gene prediction algorithms. As noted above, model integration has very wide-ranging interpretations and possible directions. Here we provide some example of how biologically motivated process models might be exploited in order to develop a richer set of models for molecular sequence and gene expression data.

Biological Criteria for Data Encapsulation

In our discussion of data integration we discussed the need to characterize and encapsulate data in such a way as to make it "context independent" and "comparable." The suitability of different encapsulation schemes often depends on the questions being asked, however a number of biological criteria are commonly invoked to establish comparability. For example, referring to portions of a proteins as a "domains" is one type of encapsulation motivated by an underlying biological hypothesis of similarity with respect to structure and function. Another primary biological model for establishing comparable and context independent units of characterization is that of homology. The concept of homology invokes a notion of "sameness" and continuity of descent (i.e. traits in different organisms that correspond to a single trait in a common ancestor; Wagner, 2001; Wagner and Stadler, 2003). period Hand in hand with homology statements is the notion that biological systems can be decomposed into sets of quasi-independent elements. That is, we recognize units of homology by the fact that such units exhibit a certain degree of independence through development or evolution. Hypotheses of homology and decomposability are widely used in genomics, either implicitly or explicitly. For example, the BLAST algorithm can be seen as generating a set of ranked hypotheses about sequence homology. In fact, statements of homology are often an absolute requirement for comparative genomic analyses.

A particularly pressing challenge if we are to achieve an integrative genomics that scales from sequences to phenotypes is to come up with useful working definitions of homology that can be applied at levels of organization above the sequence level. For example, a way of defining homology at the level of genetic subnetworks would greatly facilitate comparative functional genomic analyses across disparate genomes by focusing attention on similarities and dissimilarities among groups (subnetworks) of genes rather than individual genes. Similarly, a suitable decomposition in terms of subnetworks would be useful at the level of analysis integration. For example, given data on gene expression and protein-protein interaction we might ask whether these two types of data are telling us the same thing, not on a gene-by-gene basis, but rather at the level of sets of interacting genes (i.e. the details might be different, but the story is essentially the same). Similar challenges remain for defining homology for complex phenotypes (e.g. behavior).

Model Derived Constraints

Above we discussed some of the challenges we face when trying to integrate value schemes from different types of analyses. How, for example, might we calculate joint probabilities for outcomes at two distinct levels of biological organization or operating under distinct evolutionary models? As a concrete example, suppose we have an algorithm for producing the likely starting positions of a particular gene (annotated with value schemes) and another algorithm for producing the likely locations of regulatory elements for that gene similarly annotated. We would like to combine the information produced by these algorithms in order to come up with a combined representation of the genome in terms of gene–promoter pairs. One approach is to view the outputs of the two algorithms as random variables X and Y respectively. The value schemes associated with the algorithms give us "prior" distributions for X and Y. How might we derive the posterior distribution of interest?

We suggest that one way to facilitate such integration is to encode empirically or computationally derived biological knowledge as models that specify constraints or probabilities on the possible outcomes at a different level. For example, detailed studies of genes and promoters across might lead us to the conclusion that, generally speaking, "Regulatory elements occur within 5K base pairs of exons." This imposes a constraint on the joint distribution of X and Y. The posterior distributions of X and Y are the marginal distributions of this constrained joint distribution. To generalize, we expect the biological constraints governing the relationship between two random variables to take the form of constraints. The constraints might be (1) structural constraints which are based on the geometry of the object being found by the algorithm; (2) relational constraints in which certain set of measurements have functional relationships; and (3) abundance constraints which are constraints on how many or how few a number of some feature must occur around a particular locus. A challenge is to develop general algorithmic techniques for computing (modes of) posterior distributions under any of these types of constraints. A constraint oriented view of joint priors can lead to new algorithmic approaches. For example, joint distributions can be described more succinctly from a geometrical point of view (e.g., Allman and Rhodes, 2003; Geiger *et al.*, 2002; Kim, 2000).

Extended models of molecular evolution

Many sequence models such as HMM and generative grammar models are described in terms of "sequence production", but in reality sequences are not generated or produced in the manner described by these models. For example, HMMs describe sequence generation as a left to right production of nucleotide symbols, but this is not how the original sequences are produced by biology. In fact, the sequences have a biological generative process, namely genealogical inheritance and evolutionary change. Models of molecular evolution provide fundamental principles for inferring biological processes from static data (e.g., inference of phylogenetic trees representing genealogical relationship) and help to assess probabilistic significance of inferred patterns (e.g., probability of sequence identity between two homologous molecules). As the types of genomic data grow, there is a need to develop new models and extend currents models for new types of data and more complicated modes of evolution. Here we discuss three examples, evolutionary dynamics of short oligo tags, non–coding sequences, and gene expression.

Evolutonary dynamics of short oligomers

High–throughput technologies often use small subsequences, oligomers 〈 100 bases, as surrogate markers for larger sequences. Even under standard models of evolution for larger sequences, e.g. Poisson mutation process, the dynamics of sequence tags are not well understood. The solution to this problem will impact many techniques such as arrays that depend on the detection of short sequence motifs. As a first step problem, suppose we have a collection of sequences related to each other by a tree graph. This tree graph may represent genealogies of whole genomes or the genealogy of gene duplications within a genome. Assuming a constant rate Poisson mutation process, we wish to know the dynamics of the presence/absence of sequences tags of particular length *k*. We have studied this problem for tags of sufficient length and fixed size such that the probability of non–homologous presence is low; under this simple scenario the probability of presence of a tag of length *k* at some node in the tree is a geometric distribution with parameters dependent on the relative time–length of the left and right subtrees pending from the node. However, for even this simple problem, the solution is unknown for the ensemble of tag lengths because the different tags and tags at different length are dependent on each other.

Evolutionary dynamics of non–coding sequences

Functional non–coding sequences include promoters, enhancers, chromatin structural elements, small RNA elements, as well as introns and UTRs (e.g. Hall *et al.*, 2002; Chi *et al.*, 2003; Herbert, 2004). These kinds of elements are subject to different kinds of evolutionary forces, such as frequent insertions and deletions, sequence conversion, and transpositions (Carter and Wagner, 2002; Lynch, 2002; Hahn *et al.*, 2003). For example, Kim (2001) has studied the evolutionary dynamics of intervening sequences in between conserved sequences in the regulatory region of the *Drosophila* gene *hairy* and found evidence for neutral random change in sequence length. Functional annotation of these kinds of elements has been the focus of new research efforts (ENCODE; http://www.genome.gov/ENCODE). In particular, it would be desirable to use models of evolutionary dynamics for both algorithm development and for significance calculations. Currently, most of the algorithms and probability calculations are based on point mutations or small-order linear dependency models.

The primary work required for these studies is a systematic study of the empirical evidence. Compilations of the positional distribution of the non-coding sequences, such as frequency distribution of the length of conserved sequences, frequency distribution of intervening sequences, and their relation to the coding sequences is only beginning to be available (e.g., Stein *et al.*, 2003; Kent *et al.*, 2003; Cooper *et al.*, 2003; Hampson *et al.*, 2002; Jareborg *et al.*, 1999). Comparative genomics approaches have been useful in this context and have lead to methods for functional prediction (e.g., Levy *et al.*, 2001; Rivas *et al.*, 2001; Wasserman *et al.*, 2000). The observed empirical distributions can be used to derive heuristic significance values for predicted non-coding sequences. Statistical compilations can give heuristic assessment of significance of a given putative functionally important non-coding sequence. But, for more principled inference, an evolutionary process model will need to be developed building on the recent works studying the evolution of non-coding sequences (e.g., Kim, 2001; Rogozin *et al.*, 2002; Ohta, 1997).

Evolutionary dynamics of gene expression

Recent studies (Rifkin *et al.*, 2003) indicate that the expression level of individual genes follow an evolutionary dynamics similar to many continuous quantitative traits (e.g., body size, bristle numbers). Traditionally, quantitative trait evolution is modeled using a variety of diffusion processes (reviewed in Hansen, 1997). For example, a quantitative trait under stabilizing selection can been modeled using the Orstein-Uhlenbeck process (diffusion under a centralizing force-field; Hansen, 1997; Hansen and Martins, 1996). These kinds of models can be used to predict the expected dispersion of gene expression levels across different lineages of organisms which in turn can be used to predict the functional significance of expression level akin to the kinds of inferences made in sequence analysis. There are several open problems here: (1) methods to measure fundamental parameters of the process from data; (2) incorporation of pleiotropic and epistatic regulatory connections between different genes; and (3) extension to different kinds of continuous stochastic processes.

One of the fundamental parameters in such a model is the gene specific rate of mutational change per unit time (in the expression level). A complicating factor is that different genes have different degrees of epistatic input from other genes that affect its expression. A gene with a large number of other genes affecting its expression is also expected to show a high mutation rate because the change in any of the other genes will affect its expression. One possible approach is to estimate a characteristic degree of regulatory connectivity for a particular organism and model ensemble behavior. Finally, many varieties of diffusion processes predict a normal distribution for the trait value. Typical gene expression data sets, studied under artificial mutagenesis, show a distribution similar to a normal distribution but with a large tail indicating a mixture model of small effects (normal) and large effects (long tail) (e.g., Hughes, 2000; Rifkin *et al.*, 2000). Such mixture models can be studied by pooling the various studies and using, for example, Bayesian estimation procedures to separate the distributions. For modeling quantitative evolutionary dynamics, an important determinant is the genetic component of the linear variance-covariance structure of the traits (cf., Lande, 1979; Lande and Arnold, 1983). This is because, similar to the gradient matrix of a continuous dynamical system, the linear genetic variance-covariance matrix determines the short-term direction and magnitude of change. The main complication in the gene expression case is extreme high dimensionality and possibly complicated high-order joint distributional structure. However, from many empirical observations, it is clear that the dynamics of gene expression, whether in a developmental sense or with respect to genetic variants, do not fully span the possible dimensions, in fact all evidence suggests far less (e.g., Rifkin *et al.*, 2000). Thus, one possible approach is to reduce dimensionality by estimating the "biologically relevant", i.e., noise-free, dimensions of the gene expression space and then computing the variance-covariance structure with respect to some reasonable basis vector of the reduced dimensions.

Conclusions

In the introduction to this review we noted that many genomic approaches share at least one major feature - high throughput systematic collection of data without explicit predefined hypotheses. While such efforts have been critiqued as being "fishing expeditions" there is a real strength in leveraging large-scale data both for specific problems as well as for new problems. In particular, as our knowledge of the organism becomes more complete and complex, we expect that the nature of the questions we pose will change. Already, in the field of gene regulation we have moved from the concept of "master regulatory genes" to gene regulatory cascades. Numerous analyses of functional genomic data suggest that co-regulatory dynamics tend to be both complex and context dependent. Thus the question of interest may not be whether any given gene interacts with any other gene (the answer may be both "yes" and "no" depending on context), but rather "what is the overall architecture of genetic regulatory interactions which drives the dynamics of the system?"

Our science moves beyond the descriptive, to the "integrative", when we begin to use genomic data to test or derive specific biological hypotheses. We are at a point where we can begin to undertake such model driven approaches to genomics. However, it is our view that to do so effectively the field will have to address many of the issues raised in the previous pages. Most critical amongst these challenges is developing a unified theoretical foundation for biological data analysis and knowledge representation. In the end, all our scientific knowledge is model based or theory based. Even the proper interpretation of the movement of a nucleotide sequence on an agarose gel requires a model theoretic understanding of statistical mechanics. The main difference between such measurements and genome scale measurements is that we already have a well-defined model understanding of the movements of charged molecules in an electrical field. The future waits for a similar model theoretic view of the whole genome and the organism.

Acknowledgements

J. Kim dedicates this article in the honor of Dr. Kang Hyen Sam on the occasion of his retirement in 2004. Dr. Kang provided the inspiration and the original encouragement for the author's adventure into computational biology twenty years ago. P. Magwene acknowledges the support of a National Science Foundation Minority Postdoctoral Fellowship.

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511.

Allman, E.S. and Rhodes, J.A. (2003). Phylogenetic Invariants for the General Markov Model of Sequence Mutation. *Math. Biosci.* 186, 113–144.

Ashburner, M., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29.

Ben-Dor, A., Pe'er, I., Shamir, R., and Sharan, R. (2001). On the complexity of positional sequencing by hybridization. *J. Comput. Biol.* 8, 361–71.

Bookstein, F. L. (1991). *Morphometric tools for landmark data.* Cambridge University Press, New York.

Carter, A.J.R. and Wagner, G.P. (2002). Evolution of functionally conserved enhancers can be accelerated in large populations: a population genetic model. *Proc. Roy Soc., Biol.* 169, 953–960.

Chi, J.T., *et al.* (2003). Genomewide view of gene silencing by small interfering RNAs. *Proc. Natl. Acad. Sci. USA* 100, 6343–6346.

Collins, F.S., Morgan, M., and Patrinos, A. (2003). The human genome project: lesions from large-scale biology. *Science* 300, 286–290.

Davidson, S.B., Crabtree, J., Brunk, B.P., Schug, J., Tannen, V., Overton, G.C., and Stoeckert, C.J.Jr. (2001). K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources. *IBM Systems Journal* 40, 512–531.

Cooper, G. M., Brudno, M., NISC Comparative Sequencing Program, Green, E.D., Batzoglou, S., and Sidow, A. (2003). Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* 13, 813–820.

Cryan, M., Goldberg, L., *et al.* (1998). Evolutionary Trees can be Learned in Polynomial Time in the Two-State General Markov Model. *IEEE Symposium on Foundations of Computer Science*, 436–445.

Deutsch, A., *et al.* (1999). Physical data independence, constraints, and optimization with universal plans. *Proc. VLDB.*

Drawid, A. and Gerstein, M. (2003). A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J. Mol. Biol.* 301, 1059–1075.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids.* Cambridge University Press.

Efrat, A., Hoffmann, F., Kriegel, K., Schultz, C., and Wenk, C. (2002). Geometric algorithms for the analysis of 2D-electrophoresis gels. *J. Comput. Biol.* 9, 299–315.

Farach, M. and Kannan, S. (1999). Efficient algorithms for inverting evolution. *JACM* 46, 437–450.

Felsenstein, J. (2003). *Inferring phylogenies.* Sinaur, Sunderland, MA.

Foth, B.J., Ralph, S.A., Tonkin, C.J., Struck, N.S., Fraunholz, M., Roos, D.S., Cowman, A.F., and McFadden, G.I. (2003). Dissecting apicoplast targeting in the malaria parasite Plasmodium falciparum. *Science* 299, 705–8.

Friedman, N. and Koller, D. (2003). Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning* 50, 95–126.

Geiger, D., Meek, C., and Sturmfels, B. (2002). On the toric algebra of graphical models. *Microsoft Research: MSR-TR*, 47.

Gentle, J E. (2003). *Random Number Generation and Monte Carlo Methods.* Springer-Verlag, New York.

Hahn, M.W., Stajich, J., and Wray, G.A. (2003). The effects of selection against spurious transcription factor binding sites. *Molecular Biology and Evolution* 20, 901–906.

Hall, I.M., *et al.* (2002). Establishment and maintenance of a heterochromatin domain. *Science* 297, 2232–2237.

Hampson, S., Kibler, D., and Baldi, P. (2002). Distribution patterns of over-represented k-mers in non-coding yeast DNA. *Bioinformatics* 18, 513–528.

Hansen, T. (1997). Stabilizing selection and the comparative analysis of adaptation. *Evolution* 51, 1341-1351.

Hansen, T. and Martins, E. (1996). Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. *Evolution* 50, 1404-1417.

Harris, M. A. *et al.* (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258-261.

Herbert, A. (2004). The four Rs of RNA-directed evolution. *Nat. Genetics* 36, 19-25.

Hillis, D.M. (2000). AIDS. Origins of HIV. *Science* 288, 1757-1759.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., *et al.* (2000). Functional discovery via a compendium of expression profiles. *Cell* 102, 109-126.

Jareborg, N., Birney, E., and Durbin, R. (1999). Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* 9, 815-824.

Kennedy, D. N., Makris, N., Herbert, M. R., Takahashi, T., and Cavness, V. S. (2002). Basic principles of MRI and morphometry studies of human brain development. *Developmental Science* 5, 268-278.

Kent, W.J., Baertsch, R. Hinrichs, A., Miller, W., and Haussler, D. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* 100, 11484-11489.

Kerr, M. K. and Churchill, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics* 2, 183-201.

Kim, J. (2000). Slicing hyperdimensional oranges: The geometry of phylogenetic estimation. *Mol. Phyl. Evol.* 17, 58-75.

Kim, J. (2001). Macroevolution of the hairy enhancer in Drosophila species. *J. Exp. Zool.* 291, 175-185.

Kim, J., Moriyama, E., Warr, C.G., Clyne, P.J.,and Carlson, J.R. (2000). Identification of multi-transmembrane proteins from genomic databases using quasi-periodic structural properties. *Bioinformatics* 16, 767-775

Kissinger, J.C., Brunk, B.P., Crabtree, J., Fraunholz, M.J., Gajria, B., Milgram, A.J., Pearson, D.S., Schug, J., Bahl, A., Diskin, S.J., Ginsburg, H., Grant, G.R., Gupta, D., Labo, P., Li, L., Mailman, M.D., McWeeney, S.K., Whetzel, P., Stoeckert, C.J., and Roos, D.S. (2002). PlasmoDB: The Plasmodium genome database. *Nature* 419, 490-492. *Also* 〈http:// PlasmoDB.org〉.

Lande, R. (1979). Quantitative genetics analysis of multivariate evolution, applied to brain:body size allometry. *Evolution* 33, 402-416.

Lande, R. and Arnold, S. (1983). The measurement of selection on correlated characters. *Evolution* 37, 1210-1226.

Lauritzen, S. (1996). *Graphical Models* Clarendon Press, Oxford.

Levy, S., Hannenhalli, S., and Workman, C. (2001). Enrichment of regulatory signals in conserved non-coding genomics sequences. *Bioinformatics* 17, 871-7.

Lynch, M. (2002). Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci. USA* 99, 6118-6123.

Mantripragada, K. K., Buckley, P. G., Diaz de Stahl, T. D., and Dumanski, J. P. (2004). Genomic microarrays in the spotlight. *Trends in Genetics* 20. 87-94.

Metzker, M.L., Mindell, D.P., Liu, X.M., Ptak, R.G., Gibbs, R.A., and Hillis, D.M. (2002). Molecular evidence of HIV-1 transmission ina criminal case. *Proc. Natl. Acad. Sci. USA* 99, 14292-14297.

Ohta, T. (1997). The meaning of near-neutrality at coding and non-coding regions. *Gene* 205, 261-7.

Rahm, E. and Bernstein, P.A. (2001). A survey of approaches to automated schema matching. *VLDB Journal* 10, 334-350.

Rifkin, S. A., Atteson, K., and Kim. J. (2000). Constraint structure analysis of gene expression. *Functional and Integrative Genomics* 1, 174-185.

Rifkin, S.A., Kim, J., and White, K.P. (2003). Evolution of gene expression in the Drosophila melanogaster subgroup. *Nat. Genetics* 33, 138-144.

Rivas, E., Klein, R.J., Jones, T.A., and Eddy S.R. (2001). Computational identification of non-coding RNAs in E. coli by comparative genomics. *Curr. Biol.* 11, 1369-1373.

Rogozin, I.B., Makarova, K.S., Natale, D.A., Spiridonov, A.N., Tatusov, R.L., Wolf, Y.I., Yin, J., and Koonin, E.V. (2002). Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Res.* 30, 4264-71.

Stein, L.D. (2003). Integrating biological databases, *Nature Reviews Genetics* 4, 337-345.

Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B., and Botstein, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc. Natl. Acad. Sci. USA* 100, 8348-53.

Wagner, G. P.(2001). *The Character Concept in Evolutionary Biology.* Academic Press, San Diego.

Wagner, G.P. and Stadler, P.F. (2003). Quasi-independence, homology and the unity of type: A topological theory of characters. *Journal of Theoretical Biology* 220, 505-527.

Wang, Y. and Guo, S. (2004). Statistical Methods for Detecting Genomic Alterations Through Array-Based Comparative Genomic Hybridization (CGH). *Frontiers in Bioscience* 9, 540-549.

Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W., and Lawrence, C.E. (2000). Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.* 26, 225-227.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Jr., Marks, J.R., and Nevins, J.R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci. USA* 98, 11462-11467.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 30, e15.

Yarrow, J. C., Feng, Y., Perlman, Z. E., Kirchhausen, T., and Mitchison, T. J. (2003). Phenotypic screening of small molecule libraries by high throughput cell imaging. *Comb. Chem. High Throughput Screen* 6, 279-286.