Trends in computational biology—2010

H Craig Mak

Interviews with leading scientists highlight several notable breakthroughs in computational biology from the past year and suggest areas where computation may drive biological discovery.

The field of computational biology encompasses a set of investigative tools as much as being a research endeavor in its own right. It is often difficult to gauge the utility and significance of a computational tool, at least until the research community has had sufficient time to explore, exploit and hone it in various applications. In an effort to identify recent notable breakthroughs in the field of computational biology, *Nature Biotechnology* surveyed leading researchers in the area, asking them to nominate papers of particular interest published in the previous year that have influenced the direction of their research. Some of the nominated papers had been featured in our pages and elsewhere; others were completely off our radar. Although we surveyed a small group of 15 scientists, the nominated papers (**Box 1**) provide a snapshot of some of the most exciting areas of current computational biology research.

All the papers featured in the following pages were nominated by at least two scientists. Our analysis not only highlights the richness of approaches and growth of the field, but also suggests that researchers of a particular type are driving much of cutting-edge computational biology (**Box 2**). Read on to find out what characterizes them and what they've been doing in the past year.

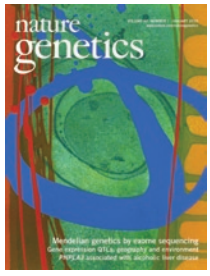
Next-generation sequence analysis

Imagine an experiment generating a billion data points every day, the equivalent of running millions of agarose gels—and taping (remember that!) the pictures into tens of thousands of laboratory notebooks, or hybridizing thousands of gene-expression microarrays. Computational biology has risen in prominence in recent years largely because of the increase in the data-generation capacity of high-throughput technology. More data create more opportunities and a more pressing need for systematic methods of analysis. And nowhere has that need been more evident than in the field of next-generation sequencing.

The latest sequencers take a week or two to generate about a billion short reads, stretches of about 50–400 bp of DNA sequenced from a longer molecule. Researchers face challenges on two levels when turning massive collections of reads into biologically meaningful information. The first set of challenges lies in processing the reads themselves: mapping them to their genomic locations, and then assembling them into longer contiguous stretches of DNA. The second set of challenges lies in interpreting large collections of reads, which may be assembled into whole genomes, to understand the functional effects of genetic variation. Thousands of genomes from humans, plants, animals and disease tissues have already been sequenced—and all are in need of better interpretation. Although algorithms, such as BLAST for searching and CLUSTALW for aligning, continue to be the workhorses of sequence analysis, several next-generation computational methods have emerged to cope with the DNA sequences captured in billions of short reads and thousands of genomes.

The advance. Two methods for *de novo* transcriptome assembly of short reads were published this year from Lior Pachter and colleagues[1] and from Aviv Regev and colleagues[2]. The transcriptome can be analyzed by sequencing cDNA reverse transcribed from RNA (RNA-Seq), but mapping and assembling the resulting reads are challenging owing to the complexities introduced by RNA splicing. The two methods are the first that robustly assemble full-length transcripts, including alternative splicing isoforms. In contrast to previous approaches, these two methods first map reads to the genome using software that takes possible splice junctions into account, thereby making assembly more manageable. Then, they apply graph-based algorithms to determine[1,2] and quantify[1] the most likely splice isoforms. The algorithms were applied to mammalian transcriptomes to follow global patterns of splicing during a developmental time course[1] and to identify novel, spliced, long, noncoding RNAs that had not been annotated by existing methods[2].

Progress toward the second challenge of genome interpretation was reported in papers that demonstrate the potential of genome sequencing for genetic analysis of human traits. The approach, pioneered by Jay Shendure at the University of Washington in Seattle, sequences the exonic regions of several genomes to identify protein-disrupting mutations linked to disease. "This study dramatically demonstrates how we can make new genetic discoveries by sequencing all the exons in a set of patients," says Steven Salzberg of the University of Maryland and a co-author with Pachter[1]. Since the publication of the first success of this strategy in January 2010 (ref. 3), several additional studies have taken a similar approach to study the genetics of human diseases.



What it means. Advances in transcript assembly from RNA-Seq data should allow alternative splicing to be studied genome-wide across

H. Craig Mak is Associate Editor, Nature Biotechnology

Steven Salzberg collaborated with Lior Pachter and Barbara Wold to develop a method for assembling short reads sequenced from cDNA into full-length spliced transcripts.

many biological conditions, such as in different tissues, over different time points and in response to genetic and chemical perturbations. In addition, RNA-Seq is now poised as a tool for discovering new RNAs that we may not have even known were transcribed from a genome. Armed with better knowledge of splicing patterns and comprehensive transcript catalogs, it should be possible to improve the annotation of genomes. "Thousands of people have accessed our software," says Salzberg, "and it is being integrated into easy-to-use graphical interfaces such as Galaxy"[4].

The flood of whole genomes and exomes should also drive sequence-analysis methods. Attending an International Cancer Genome Consortium meeting in Brisbane, Australia, in December, Debbie Marks of Harvard University noted, "We're still in the early days of whole-genome analysis…problems need to be articulated in ways that are computationally tractable." Many of the problems will require algorithmic advances, such as better *de novo* assembly of transcriptomes and genomes. However, much of the work that goes into interpreting a patient's whole genome to identify disease-causing mutations, for instance, involves filtering variants identified by sequencing against databases of known variants. As better databases should lead to improved genome analysis, it might be reasonable to expect the development and mining of biomedical databases to be a fertile source for computational advances.

1. Trapnell, C. *et al. Nat. Biotechnol* **28**, 511–515 (2010).
2. Guttman, M. *et al. Nat. Biotechnol.* **28**, 503–510 (2010).
3. Ng, S. *et al. Nat. Genet.* **42**, 30–35 (2010).
4. Goecks, J. Nekrutenko, A. & Taylor, J. *Genome Biol.* **11**, R86 (2010).

Discovery from data repositories

Electronic medical records are becoming a reality, promising lower health-care costs, improved patient treatments and, perhaps, scientific advances. As a result of incentives built into the Health Information Technology for Economic and Clinical Health (HITECH) Act passed as part of the Obama administration's health-care reform, in 10 years almost every US hospital could be using electronic medical records, up from 1.5–2% of hospitals today. "What's fun to think about," says Atul Butte of Stanford University, "is what kind of science can we derive from this?"

Electronic medical records can contain a wealth of history on physical exams and treatment regimes. Particularly amenable to automated analysis in the records are the standardized administrative billing codes used to charge for each procedure, test or clinical visit. These codes can track anything from diagnosis of coronary artery disease to the procedure for inserting a stent to keep blood vessels open. Several hospitals, with Vanderbilt University (Nashville, TN, USA) among the leaders, are pairing electronic medical records with the collection of tissue samples from every patient treated. These resources represent an unprecedented source of data on the genetic and physiological state of people linked to standardized, computable records of their phenome, or the set of all phenotypes including disease diagnosis and responses to treatment.

Atul Butte: "Ninety-nine percent of the work is not in software engineering or coding, it's in coming up with the right kind of question:…[one that] no one even realizes we can ask today."

The advance. Last year, Joshua Denny and colleagues at Vanderbilt University published the first study that demonstrates the feasibility of associating genetic modifications with data on phenotypic traits mined from electronic medical records[1]. The approach, which they called PheWAS (for phenome-wide association scans), is akin to the genome-wide association studies (GWAS) widely used today to find single-nucleotide polymorphisms (SNPs) that are genetically linked in a population to a particular disease trait—except that PheWAS is GWAS in reverse. GWAS associates genotypes with a given phenotype, such as height or a genetic disease. In contrast, PheWAS attempts to determine the range of clinical phenotypes associated with a given genotype.

The Vanderbilt group analyzed the medical records of ~6,000 patients who had been tested to see whether they carried a total of five SNPs previously associated with seven diseases (coronary artery disease, carotid artery stenosis, atrial fibrillation, multiple sclerosis, lupus, rheumatoid arthritis and Crohn's disease). To identify patient phenotypes in an automated fashion, they used billing codes in the electronic medical records to group patients into 'case' and 'control' populations for 776 phenotypes. Finally, a Chi-squared statistical test was used to evaluate whether patients harboring a specific SNP also tended to display a particular phenotype. The authors noted that, although there are many statistical challenges with this kind of analysis and there is much room for improvement of their method, four of the seven previously known disease-gene associations could be replicated, and several potential associations with other diseases were identified but not rigorously validated. These results highlight the possibility that novel biological discoveries might be made using this approach.

What it means. "Everyone wishes they could do this kind of study," remarks Butte, "but it represents a multimillion dollar investment. Vanderbilt is leading the way." Several other hospitals are making similar investments. The Mayo Clinic, for example, is coupling specimens collected from 20,000 patients with electronic medical records and other data gathered and standardized across the hospital system. "There has always been a question about whether electronic medical records would be of sufficient quality to allow genetic discovery," says Russ Altman, also at Stanford. "This paper sets the stage for widespread use of electronic records for genomic discovery."

More generally, the case of electronic medical records illustrates the potential value locked within unique biomedical databases and the challenges of realizing that value. For instance, a paper describing the PubChem BioAssay database[2] has caught the attention of several survey respondents. PubChem is the repository for small-molecule screening data generated by several NIH programs, and it receives similar data from many other organizations. "PubChem will become a key technology, in a manner similar to how freely available sequence databases in the 1990s enabled a generation

of computer-literate biologists to change the way biology is done," says Iain Wallace, a postdoctoral fellow in Gary Bader's group at the University of Toronto, which has been active in the development of databases of protein interactions. PubChem, which is funded by the NIH, brings to academics data that until now have been accessible primarily only to those in deep-pocketed pharmaceutical companies.

In the case of PubChem or Vanderbilt's electronic medical record database, careful statistical analyses will be required to robustly analyze these potential treasure troves of information. But rather than the algorithmic advances typically pursued in computational biology, according to Butte, "Ninety-nine percent of the work is not in software engineering or coding; it's in coming up with the right kind of question: given this data set, what question are we newly able to ask that everyone would love to know the answer to, but no one even realizes we can ask today?" Exposure is key, says Butte, "What I would love to see is a computational person going to surgical grand rounds at a hospital to figure out what the unsolved questions are, hearing about this tumor that spreads like crazy and saying, 'I can solve this problem computationally.' That would be the ideal." Unlike problems requiring clever new algorithms or massive clusters of computers, increasing exposure may be a particularly manageable challenge facing the field.

1. Denny J.C. *et al. Bioinformatics* **26**, 1205–1210 (2010).
2. Wang, Y. *et al. Nucleic Acids Res.* **38** database issue, D255–D266 (2010).

Learning to see

Why have computer scientists long endeavored to create software capable of accomplishing tasks humans can already do? In the case of biological research, one advantage of computational analysis is automation and fidelity. Whereas a trained person can look at one confocal microscope image and readily identify where a fluorescently labeled protein is localized in the cell, that person cannot hope to analyze the millions of images that can be gathered with automated technology. And even if several people were enlisted to the task, each may interpret the same image in different ways. This problem provides an apt introduction to machine learning, a technology that is finding success in biology.

In machine learning, computer programs are trained to pick out patterns, which may be predefined by human supervisors or learned by the program directly from data. Such 'unsupervised' machine-learning tasks are often the hardest, in part because there are many possibilities for the computer to consider. Notably, many machine-learning tasks in disparate problem domains can be articulated using a common set of concepts. In this way, techniques developed for one problem, say mining data from text, can inspire solutions to other problems.

Robert Murphy thinks that when computer science and biology come together "inside one person's head, that is a much more efficient process."

The advance. Robert Murphy and colleagues[1,2] at Carnegie Mellon University devised machine-learning algorithms that could accurately classify whether a pattern of fluorescent staining represents localization to one subcellular organelle or to a mixture of locations. Moreover, this 'pattern unmixing' can be done in an unsupervised way, without introducing bias from a human who predefines the categories. The need for this method is supported by studies in yeast in which up to a third of all fluorescently tagged proteins appeared to localize to several places in the cell.

The key to the approach is to segment an image into objects or shapes with quantifiable features. Then a pattern of objects can be defined as the probability that certain objects are found together. The best-performing algorithm identified patterns of objects using a technique called latent Dirichlet allocation, which has been successfully used to identify patterns of words representing conceptual topics from text documents. By analogy, visual objects representing the nucleus or Golgi apparatus are 'words' in an image, and patterns of protein localization that characterize the content of an image correspond to sets of words that co-occur in documents and define the topics in the text.

What it means. "This represents the first step toward a new way of thinking about interpreting images that is generative rather than descriptive," says Murphy. Whereas a descriptive approach may take an image of a cell expressing fluorescently tagged protein and tell you that the protein is in the nucleus, a generative approach builds a model that can produce images that look like other images, and in the process of building that model (that is, determining the parameters of the model), you learn about what characterizes a pattern in a way that is meaningful across a variety of situations. For instance, a drug in a screening assay may cause a protein to partially redistribute from one subcellular location to another, but given that organelles may look different in different cell types, without Murphy's approach, if the same screen is done on a different cell type, it is difficult to know that the same process is occurring. Machine learning has been previously applied to biology, but recent increases in the data-generation capacity of technology suggest that these kinds of approaches may play a growing role in biological discovery in the future.

Does this mean that more collaboration needs to occur between biologists and computer scientists classically trained in machine learning? Not necessarily, according the Murphy. "That's been going on for a long time already. In fact, there is a group of people who are knowledgeable in many of these different domains. There are people who in general may not push the frontier of computer science, but who use state-of-the-art techniques, and in some cases do end up pushing frontiers and identifying new problems that others in the field can then solve." The role of computational biologists is to be able to straddle domains. Murphy continues: "When the field started, it often grew by adventitious 'collisions' between computer scientists and biologists—over lunch, at a faculty meeting. That is a very inefficient way of moving forward. When those collisions can happen inside one person's head, that is a much more efficient process."

1. Coelho, L.P. *et al. Bioinformatics* **26**, i7–i12 (2010).
2. Peng, T. *et al. Proc. Natl. Acad. Sci. USA* **107**, 2944–2949 (2010).

CompBio 2.0

Businesses and broad segments of society have recently embraced decentralized mechanisms of information processing based on interactions among large groups of people. This advance in computing has not relied on new algorithms or clever data structures in the traditional sense

Box 1 Survey results

Thirty-three papers were nominated covering genomics, imaging, databases and data sets, protein-structure prediction, synthetic biology, genetics, antibody screening, systems biology and pharmacology. The 24 papers not discussed in the article are listed below.

Akavia, U.D. *et al.* An integrated approach to uncover drivers of cancer. *Cell* **143**, 1005–1017 (2010).

Ashley, E.A. *et al.* Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525–1535 (2010).

Bandyopadhyay, S. *et al.* A human MAP kinase interactome. *Nat. Methods* **7**, 801–805 (2010).

Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465**, 53–59 (2010).

Berger, S.I., Ma'ayan, A. & Iyengar, R. Systems pharmacology of arrhythmias. *Sci. Signal.* **3**, ra30 (2010).

Carro, M.S. *et al.* The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318–325 (2010).

Coulet, A., Shah, N.H., Garten, Y., Musen, M. & Altman, R.B. Using text to build semantic networks for pharmacogenomics. *J. Biomed. Inform.* **43**, 1009–1019 (2010).

Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).

Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).

Gibson, D.G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52–56 (2010).

Lestas, I., Vinnicombe, G. & Paulsson, J. Fundamental limits on the suppression of molecular fluctuations. *Nature* **467**, 174–178 (2010).

Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

McGary, K.L. *et al.* Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc. Natl. Acad. Sci. USA* **107**, 6544–6549 (2010).

McLean, C.Y. *et al.* GREAT improves functional interpretation of *cis*-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).

Mungall, C.J. *et al.* Integrating phenotype ontologies across multiple species. *Genome Biol.* **11**, R2 (2010).

Pandey, G. *et al.* An integrative multi-network and multi-classifier approach to predict genetic interactions. *PLOS Comput. Biol.* **6**, e1000928 (2010).

Patel, C.J., Bhattacharya, J. & Butte, A.J. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS ONE* **5**, e10746 (2010).

Peng, H., Ruan, Z., Long, F., Simpson, J.H. & Myers, E.W. V3D enables real-time 3D visualization and quantitative analysis of large-scale biological image data sets. *Nat. Biotechnol.* **28**, 348–353 (2010).

Ravasi, T. *et al.* An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**, 744–752 (2010).

Reddy, S.T. *et al.* Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat. Biotechnol.* **28**, 965–969 (2010).

Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).

Shaw, D.E. *et al.* Atomic-level characterization of the structural dynamics of proteins. *Science* **330**, 341–346 (2010).

Voelz, V.A., Bowman, G.R., Beauchamp, K. & Pande, V.S. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1–39). *J. Am. Chem. Soc.* **132**, 1526–1528 (2010).

Zhang, C. *et al.* Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science* **329**, 439–443 (2010).

of computational breakthrough but rather has been fueled by new communication media and tools, typically accessed online through websites and mobile devices. The social network Facebook (http://www.facebook.com/) has evolved from a platform solely devoted to keeping up with friends into a business that will generate an estimated $1.1 billion dollars in 2010. And notably, there is still room for specialized social networks, such as Jumo (http://www.jumo.com/), a fledgling effort geared toward philanthropy.

There have been some successes in adopting distributed computing tools and social networks into biological research. Social bookmarking tools such as del.icio.us (http://www.delicious.com/), CiteULike (http://www.citeulike.org) or Connotea (which was created by Nature Publishing Group; http://www.connotea.org/) allow users to tag and share papers and have existed for several years. Patients with the same medical conditions can connect with one another on social networks run by companies like PatientsLikeMe (http://www.patientslikeme.com/) or CureTogether (http://curetogether.com/). And for ~10 years, the Folding@home project (http://folding.stanford.edu/) has been leveraging participants' desktop computers or gaming consoles to study protein folding. But how else can computational paradigms that have permeated broader society be harnessed to drive scientific discovery?

The advance. Two papers[1,2] identified by our survey respondents highlight the potential impact of 'nontraditional' computing advances. The first is FoldIt, a multiplayer online game for predicting protein structures. David Baker and colleagues at the University of Washington in Seattle created a Web-based graphical interface that allowed players to manipulate a protein structure as if they were solving a visual puzzle[1]. This harnessed humans' spatial reasoning skills to improve computational predictions of the most likely protein conformation. Players competed against one another and were ranked on a scoreboard. When protein structures derived by FoldIt players were compared against structures predicted by a traditional computational approach, FoldIt predictions were as good or better in seven of ten test cases.


Web-Based, Participant-Driven Studies Yield Novel Genetic Associations for Common Traits

In another approach[2], researchers at Columbia University and Stanford collaborated with the consumer genetics testing company 23andMe (http://www.23andme.com/) to identify associations between genetic markers and human traits. What's notable in this study is that trait data were collected through Web-based surveys completed by consumers whose genetics had been analyzed by the company. Twenty-two traits, ranging from hair and eye color to the ability to smell the urinary metabolites of asparagus, were studied in nearly 10,000 people of northern European ancestry. The study identified single-nucleotide polymorphisms known to be associated with six of the traits, as well as novel associations with four traits. Pitfalls may be associated with approaches such as this, however, including recent concerns raised over the concordance between results of genetic tests conducted by different direct-to-consumer companies.

What it means. "Social networks need to appeal to people's selfish side," says Andrew Su, Associate Director of Bioinformatics at the Genomics Institute of the Novartis Research Foundation (San Diego, CA, USA), whose group has developed collaborative scientific tools for use publicly and within Novartis (Basel). "There needs to be some personal value derived from social networking; otherwise, where's the motivation to participate?" Arguably, the gaming aspect of FoldIt appealed to participants' competitive juices. In the 23andMe study, participants had already received their genetic data from the company, and the Web surveys served to increase the value of those data. The key, then, is to

Box 2 Cross-functional individuals

In the course of compiling this survey, several investigators remarked that it tends to be easier for computer scientists to learn biology than for biologists to learn computer science. Even so, it is hard to believe that learning the central dogma and the Krebs cycle will enable your typical programmer-turned-computational-biologist to stumble upon a project that yields important novel biological insights. So what characterizes successful computational biologists?

George Church, whose laboratory at Harvard Medical School (Cambridge, MA, USA) has a history of producing bleeding-edge research in many cross-disciplinary domains, including computational biology, says, "Individuals in my lab tend to be curious and somewhat dissatisfied with the way things are. They are comfortable in two domains simultaneously. This has allowed us to go after problems in the space between traditional research projects." A former Church lab member, Greg Porreca, articulates this idea further: "I've found that many advances in computational biology start with simple solutions written by cross-functional individuals to accomplish simple tasks. Bigger problems are hard to address with those rudimentary algorithms, so folks with classical training in computer science step in and devise highly optimized solutions that are faster and more flexible."

An overarching theme that also emerges from this survey suggests that tools for computational analyses permeate biological research according to three stages: first, a cross-functional individual sees a problem and devises a solution good enough to demonstrate the feasibility of a type of analysis; second, robust tools are created, often utilizing the specialized knowledge of formally trained computer scientists; and third, the tools reach biologists focused on understanding specific phenomena, who incorporate the tools into everyday use. These stages echo existing broader literature on disruptive innovations[1] and technology-adoption life cycles[2,3], which may suggest how breakthroughs in computational biology can be nurtured.

1. Christiansen, C.M. & Bower, J.L. Disruptive technologies: catching the wave. *Harvard Business Review* (1995).
2. Moore, G.A. *Crossing the Chasm: Marketing and Selling High-Tech Products to Mainstream Customers* (HarperBusiness, 1999).
3. Rogers, E.M. *Diffusion of Innovations* (Free Press, 2003).

discover how to incentivize individuals in such a way that they support scientific discovery. One possibility is being tested by InnoCentive (partnering with Nature Publishing Group; http://www.innocentive.com/), which allows participants to pose scientific problems and offer cash prizes to other participants who provide a solution.

As in real life, different types of social interactions may justify different social networks, such as LinkedIn (http://www.linkedin.com/) for professional networking, which has thrived, even in the shadow of more general-purpose larger networks like Facebook. Several research-oriented efforts have been started, such as Sage Bionetworks (http://sagebase.org/), whose CEO, Stephen Friend, predicted earlier this year the coming obsolescence of "hunter-gatherer approaches, where large groups collect massive clinical and genomic information and expect that they as the data generator will be the data analyzer" (http://www.xconomy.com/national/2010/01/06/five-biotechnologies-that-will-fade-away-this-decade/). The two studies discussed above demonstrate successful applications of alternative paradigms for data analysis and data generation. When recruiting expertise to create these kinds of platforms, says Su, "it's hard to find people who have really traversed both computer science and biology. Discovery-oriented computational biologists with experience working on collaborative projects involving experimental scientists are particularly valuable."

1. Cooper, S. *Nature* **466**, 756–760 (2010).
2. Eriksson, N. *PLoS Genet.* **6**, e1000993 (2010).