DOE-NHGRI Informatics Workshop
4/2-3/98
Dulles Hilton

Crude notes by Jeff Schloss

Large-scale sequencing LaDeana Hillier

Three levels of informatics needs:

1) Sequence Production
   - Data tracking
     - -customized. Much support should be in the large-scale centers, but need it outside centers, too
     - -external access required
   - Physical Mapping
     - -automated band calling. Research is needed
     - -map assembly (algorithms/interfaces)
     - -map publication/dissemination
   - Data collection/analysis need larger variety of tools. Developers need to use the large amount of data that's available. Must be clearly demonstrated to be statistically significant before putting into use.
     - -lane tracking
     - -image analysis
     - -basecalling/confidence values
     - -algorithms and interfaces
   - Data Processing
   - Finishing (interfaces and algorithm development are both needed)
   - (Wash U. is sequencing at 120 Mb/yr but finishing at 60)
     - -editing
     - -problem visualization, identification, resolution
   - Technology development
     - -re-arraying
     - -plaque picking
     - -data collection
   - Systems
     - -high availability files/applications
     - -networking
2) Sequence Analysis
   - LIMS
   - Gene Prediction Aids
     - -EST/protein to genomic sequence
     - -multiple alignment tools
   - Gene Prediction Tools
     - -full clone sequence
     - -mouse sequence
     - -experimentally verifies training/test sets for development and comparison. Need support to develop this, or we'll never be able to validate tools.
   - Gene Identification
     - -gene naming conventions
     - -integration of genetic and physical maps
     - -protein domain identification tools
     - -graphical representation/integration
   - Annotation representation
     - -uniform annotation at separate sites (and uniform views of annotation)
     - -public data mining/data analysis
     - -graphical representation
3) Public Databases
   - Education
     - -users
     - -developers and bioinformaticists
     - -phase I and II data
     - -central public repository for external tools (maybe just the top 10 for each kind of analysis)
     - -standardization
   - Submission/update mechanisms
     - -sequence/mapping data (human vs. other) Need to enforce that certain fields must be filled for submission to be accepted.

-accession number/identifier policy (stability)
Stable Identifiers
-daily updates – dead gi, live gi – tools
(more)
Data Access
-sharing between public databases
-standardized , well-documented centralized formats
-libraries of routines callable by JAVA and Perl
-database entries returned could contain hotlinks
-stitch together genomic entries in a parsable way
-careful references on annotation (done by third party? Primary sequence data "owned" by submitter)
-integrated map/sequence views

Temple Smith: some groups, particularly in companies, have already solved data tracking. Wouldn't it be more effective to just obtain their software by contract rather than developing it independently everywhere?
Hillier: We do some transfer when we can; talk with them a lot. Some issues are custom.
Lipman: I've also visited those companies and none of them are satisfied with their ability to do this. Large-scale centers that still exist have solved these problems rather well. What's left are the problems that are unique to each of the groups.
Temple Smith: Even if can't buy existing systems from industry, maybe genome agencies could contract out the building of particular systems, giving the business to competent practitioners and getting the job done right.
Kucherlapati: there are some things that many people want to be able to do. Users may not be perfectly happy with every aspect of the software (as is true of all commercial software) but overall, it satisfies their needs.

Takashi Gojobori, DNA Data Base of Japan (DDBJ)

Research activities:
Developing genome sequence broker
Comparing genome structures using all avail organism sequences
Developing tools for comparing/identifying variable regions (silent or potentially functional), in addition to the more conventional comparisons of conserved regions.
Important informatics topics for the future:
-genomic diversity in humans
interface for DNA chips
measurement of mutation rate based on sequence information
-comparative genomics: evolution of genomic structure
intelligent data mining approaches. Need more efficient tools because current calculations take too long.
-genomic engineering
understanding knock-out experiment results; doing combinatorics of knock-outs. (JAS: how is informatics going to solve this problem?)
-cDNA expression profile database
can develop infrastructure around human data and build around that the ability to compare with other organisms.

Langley asked about inter-db policy issues.
TG: Can have same policy of data release. Some is marked hold until publish and we want to honor that. Patent issues complicate how we react to certain inquiries. In Japan, important date is the date of patent application filing. In U.S. it is discovery date.
Lipman: in US we release data on date when paper is published. There have been cases where the journal was available in US but not in Japan, so there were different release dates. Also policies are different for release of data from large scale production centers in US versus Japan.
Branscomb: In US policy there's big difference between instantaneous release from large centers (and no patenting) versus the policy for all the rest of the sequence producers who can hold and patent data. Think this needs to be addressed, but not at this meeting.

Chakravarti asked how much novel software is being produced in Japan.
TG: for db construction, Japan constructed independently. For genetic variation (e.g., MHC serotyping), Japan constructed it themselves. It is a small effort, even combining academic and private sectors.

Anne Spence – Medical Genetics

Perspective is from talking to parents of autistic child. User of the data. Will present through series of stories.
1. Chipping on stone. How can we profit from our mistakes? Db was going to be genes/locations (a map) Meetings were the same as what we're having today. OMIM has come a long way to representing the data the way we need it. GDB was filling important function so now there's a hole. Model org dbs are starting to incorporate population data and

that's great. But there are 3 basic problems users struggle with.

a. genetics vs. computers. Just the facts vs. creativity. How much resources to put into capturing the data versus facilitating creative use of the data. Constant struggle. Let's not forget to get the data in.
b. Seeking the wave. Nobody appreciates how much data will hit us (as we didn't do before).
c. Right vs. might. Accuracy and completeness.

2. Chipping on stone, the sequel. ADHD and reported association with DRD4. Some reasonable biology may supplement the statistics (re. receptors in neurons). Took us 3 hours for Moyzis and I (who basically understand the resources) to get the information from the db on DRD4. The annotation of the ORF isn't in the db. If go to the citation, find the wrong one. Most of the primers given in the record aren't right, either.

So here's a polymorphic locus reportedly associated with ADHD, schizophrenia, novelty seeking, and several other important conditions, and the record is in terrible condition, plus there's no population data because that's not the design of the database. So here's an example of a simple query that it took 3 hours to get a simple answer. Can't expect us to have your perspective on the world; need to serve our perspective.

Links among databases are absolutely essential. And they have to be accurate.
Regular, rapid updates.
Accurate/annotated information. Need relational hooks.
Population data linked to the sequence.

Goodman: these are not hard things to solve and they're the same questions all biologists ask. Part of the problem in solving them is reflected in the organization of this meeting. None of the breakout groups address this.
Anne: That's why I organized my talk the way I did.
Collins: Agree that this flavor is what this meeting is about. How might GDB have helped? What holes need to be filled?
Spence: GDB was a bridge between OMIM's clinical detail and the sequence data. You have to enter Genbank knowing the genetic detail. More than that, it gave broader context.
Valle: Just want to comment that this aspect of the databases is going to have to be very user friendly. Anne is pretty sophisticated about this and she couldn't find the data.
Cottingham: After the 3 hour ordeal you obtained useful information. How can others in the world benefit from your having learned this. Should there be a requirement on the db that a user can report back to the db to capture this, and make it available to others?
Spence: there's no incentive to do this. I'm not going to sit down and try to figure out how to get this into the db.
Lipman: I run into the same thing. I send in an update in an e-mail message; don't worry about formatting. We get a lot of these, but I'm sure most people don't bother. We'll try to follow it up. We also agree that there should be an up-to-date record for genes about which we know something. Working on that. Disagree strongly with Nat that this is easy. We'll never keep up completely; users will always have to do some work.
Ostell: GenBank is a user driven archive. OMIM is one person's review of the information. Anne points out there's a space between these. Now we have a new project to fill in this space. Answer doesn't exist yet, but it will.
Gelbart: Does ownership get in way of updating?
Lipman: That's one reason we have idea of reference sequence; that becomes less of an issue.
Chakravarti: user community is heterogeneous. A single solution won't be useful for everybody.
Spence: I'm a pragmatist. Know of no problem that's reached consensus in academic setting. Early databases spent too much time trying to satisfy this impossible thing. Listen when there are a few simple basic changes suggested. But don't want paralysis due to trying to reach impossible consensus.

Debbie Nickerson – Genetic Variation

There's no way to get from current databases information on the relationship between what's known about variation and the function of that variation. There's no way to know, from what's in the database, if a difference I see in a sequence is a mistake or a variant (in part because there's no indication of the quality of the sequencing that was done).
The kind of variation information that's needed is:

- Type (substitution, insertion/deletion)
- Location (how do we represent this information?)
- How was it discovered?
- Mode (inherited, acquired/induced in vitro?)
- Frequency (population, haplotype)
- Method of genotyping
- Phenotype/function

There needs to be a way to add this kind of information to existing records.

Cottingham: how to you propose to represent complex phenotypes?
Nickerson: need a simple format that people can use to update existing records (third-party annotation).

Gelbart: vocabularies are hard, but absolutely essential for complex phenotypes. Need lots of work on controlled vocabularies.
Collins: new set of rules were just published for describing sequence variations, including which bases were inserted/deleted in a repeat where there is ambiguity.
Nickerson: not enough. Doesn't address non-coding regions. It is a start.
Ostell: Beaudet paper purposely decided not to address coordinates because there was no reference sequence. It was decided to do it relative to what is known. Better numbering system will come with the data. Even though any reference sequence will be arbitrary, it gives numbering system. On issue of variation, users say they don't want to submit the whole sequence when they find a SNP -- just want to submit the difference. But if you submit the whole thing others will know what you sampled. So you need to record the data and lay on top of that the interpretation. The data are less likely to change than the interpretation. Might want to change the reference sequence, and with these new data, could do this. On the issue of function, this is the hardest. Agree strongly with Gelbart that throwing open to third party annotation is risky. Agree that you need a place to record observations, but need to figure out effective way to do this.
Gelbart: Different models. Dangerous to have 3rd party annotation on generalized db like GenBank. That's why you need specialized databases to normalize information to the system. Then need to figure out hierarchies that allow users of generalized database to get to the specialized databases.

Roger Brent – Functional Genomics

- My talk is based on an analogy by a speaker at a conference, that knowing sequence is like knowing characters in a Shakespeare play, and knowing gene expression patterns allows you to know the plot. This isn't a bad analogy. Except you need further kinds of information, not just expression data, to know the plot.
- Factories will produce categorizable data (controlled vocabularies).
- Need to deal with fuzzier data. Current computer science doesn't do well at layering inferences on each other. The narrative that can be generated from queries should be at least as deep as the understanding of most molecular biologists. Most informaticists are not at this level.
- Biologists need a lot of help from computer scientists to be able to string together hard facts and soft facts and make inferences.

Goodman: All data analyses begin by appearing to require huge amounts of human inference. But it turns out often not to be true. (You've made the problem much harder than it is and done a disservice.)
Brent: I've already been able to use it to determine some kinds of data that I believe will essential to have in order to make the inferences that I think need to be made.
Cottingham: We're getting at heart of why informatics has failed, which is problems of communications across disciplines.
Gelbart: Need to capture in dbs a sense of the biological reality versus the experimental data. You've also brought up for the first time at this meeting the need to map protein interactions and networks, which we have to figure out how to map on to linear data consisting of maps/sequence.
Temple Smith: the model for our existing dbs is a scientist working with focus on a particular protein. This is not the vision we want for the future. Person working the old way doesn't want resources taken away from existing databases, to build the ones of the future. And that person is a much more narrow biologist than the bioinformaticist who DOES have a vision of what the future database should look like.

Ranier Fuchs – Industry perspective

There is no single industry perspective because industry is diverse. I'm speaking from pharmaceutical perspective.
Industry cares about bioinformatics for
Drug development process (more tailored to individual patient – pharmacogenetics/genomics)
Identifying biological targets. Need more targets to keep pipeline filled. Also need to make the process more efficient (from 100 starting points, only 10 make it to end and only 3 result in significant profit). Bioinformatics can help to validate targets and this is crucial (and goes much beyond target identification).

Challenges are in data analysis and in knowledge discovery. Database itself has no value. NEED STANDARDS and this must come from the agencies. Also need training.
1. Data types. Need to move away from focus on sequence. How to represent
gene expression (Need much more experience, though, before we'll know how to represent these data), molecular interactions, gene regulation, genetic variation (polymorphisms, splice variants, post-translational modification). We're at the point now where we were with sequence years ago with sequence.
2. Data analysis. More precise prediction of gene function. Saying it is a kinase or transmembrane protein is much to simple a construct.
3. Modeling. Collection of facts to inferences to predictions. Virtual cell. To do this, need qualitative and quantitative descriptions of molecular interactions.
4. Tools for the rest of us. Tools must be made not only available but usable. Commercial-quality code is needed, as is support. There are people who could help create expert systems. Agencies need to make this happen.