

Annotation of the Human Genome

Justin L. Cotney

With the release of a working draft of the human genome sequence comes the next logical step of annotating this information (*1*). The draft sequence is simply massive amounts of text consisting of A, C, T, or G at a local level. Just as in a book, reading individual letters on page will yield very little interpretable information, but taken together as a whole gives a complete story. Continuing with the example of a novel, for one to truly understand its meaning, the letters must be compiled into words, sentences, paragraphs, and so on. This is the problem we are faced with in the human genome sequence. We have the majority of the letters needed to compile a complete novel, but we do not know everything about the language (such as the syntax, common word usage, and advanced sentence construction). We must be able to translate the raw text into words (codons), sentences (complete genes including exons and introns), paragraphs (splice variants and regions with no visible coding information), chapters (chromosomes), and finally a complete novel (entire genome).

To achieve such a task we must first understand the language of the genome. The draft sequence provides much of the chromosome location information but does not completely document repetitive regions. This provides a hurdle to identify the total number of genes, especially at the ribosomal DNA array, and detect variation amongst these repeats. There is extensive knowledge of the genetic code and how it specifies amino acids. There is also quite a bit of knowledge of how often particular amino acids are specified and used in genes from particular organisms. We know simple gene construction from prokaryotic organisms and simple eukaryotes, such as a start codon,

~100 amino acids, and a stop codon. However as we move to analyze more complex eukaryotic genomes the genes they encode become more disparate in the overall sequence as well as being broken into many exons. In higher vertebrates, genes are known to span as many as several hundred kilobases and contain numerous exons. Annotation of such genes provides a daunting task. Not only are the genes broken up into many pieces and spread over large distances but they also include regions that may be transcribed but not translated.

A large portion of the effort to annotate the human genome has been put forth by *ab initio* gene prediction using various algorithms. Programs including GeneScan, GeneMark, and Genie have been developed to aid researchers in gene identification (2). While these programs have made great strides in sifting through the genome, they still fall short when it comes to identifying large complex genes as well as exceedingly short open reading frames. The problems lie in the ability of the programs to accurately identify exon boundaries as well as not being able to identify genes outside of predefined minima or maxima of open reading frame length. Additionally these programs also depend on an outside source of information to confirm a gene hit, such as cDNA identification, which can severely hamper ones ability to discover new genes (2-4). So in order to improve the ability of programs to predict genes within a genome a consensus must be reached for the description of a gene in terms that can be readily adapted to programming languages. The programs must also be able to make predictions about genes without relying on previously existing outside information.

Here I propose developing a new approach to gene finding in order to completely annotate the human genome. I do not suggest completely scraping the previously

developed algorithms, but taking the strong points of each and applying them in an integrated approach. First I need to define what a gene is composed of in this particular study. A gene in its complete sense will be composed of, most importantly, the region of DNA that is responsible for directly coding for an amino acid sequence or an RNA molecule that serves a functional role such as ribosomal RNA. In addition to the coding sequence, any internal non-coding sequence such as introns must be included. Extending from these regions we must then include any promoter regions and transcribed but untranslated regions of sequence. The approach I would like to take would make several passes over the entire genome with different search parameters at each successive level. In essence I would like to find a gene at its most basic level then expand outward in the complexity of the annotated gene.

This process would require the use of previously devised algorithms for defining exon and introns boundaries, such as those used by Genotator (5). In addition, I would like to expand where the program looks for coding sequence. Previously once a gene was identified in a particular region of DNA, programs would not check for additional genes within the same sequence nor on the opposite strand. We now know that genes can overlap and be encoded in opposite directions to previously identified genes (5). This opens the possibility for overlapping genes as well as coding sequences in both directions from the same stretch of DNA. I propose not to limit the gene finding by any of these previous constraints and allow prediction in all six reading frames for any given region of DNA. Additionally I would like to expand the limits of open reading frames. In the first attempts to discover genes in the yeast genome, an arbitrary cut off of 100 amino acids as the minimum open reading frame. In these studies it was found that over 2000 open

reading frames were shorter than this predefined length, representing nearly a quarter of those identified in yeast (6-9). In my approach I propose that all identified reading frames be tested for ability to form a globular structure through proteomics comparisons using resources from PDB, PFAM, PROSITE and SWISS-PROT (10-13) databases and then included or excluded from analysis based on those results. This would comprise the first pass of the analysis: identification of all possible open reading frames, including short, long, and exon containing, and prediction of those that may or may not be actual protein coding genes. In addition to the protein coding genes, at this stage any RNA only coding genes, such as ribosomal and transfer RNAs, would be identified. This is a fairly straightforward task due to the limited number of genes that encode for such molecules, but it will be difficult to predict new genes that might only be expressed at this level.

The second phase of analysis would consist of taking those predicted genes and more thoroughly defining splice sites for exons as well as untranslated, but transcribed regions at both the 5' and 3' termini of the predicted gene. This is a more complicated task than the first task, but by focusing only on those regions that are believed to contain coding information the time to process this data will be dramatically reduced. This analysis would consist of algorithms designed to find starts of transcription based primarily on identification of the closest conserved promoter regions independent of sequence length. The determination of the 3' end of the transcript or the end of the gene is a more difficult task. Since most vertebrate transcripts have a poly-A tail it is difficult to predict the actual stop of transcription. A more simple approach and possibly more relevant one would be to identify the poly-A signal site and use this as a crude 3' terminus of the gene.

Once this analysis had been completed for each of the possible open reading frames many could be excluded from further analysis by whether they had detectable promoter sites and 3' terminus processing signals. The last phase of annotation would consist of analysis of regions near those genes that had made it through the first two passes to discern what transcription factor binding sites might be present. Analysis of both stands within 50 kb of either terminus of predicted genes with the TRANSFAC database would allow a basic prediction of how each protein might be expressed and under what controls (14). The analysis could then extend to find enhancer regions and possible sites of epigenetic control (15) if such a level of annotation is desired at this time.

While this may seem as an oversimplified presentation of the approach, I think many of the technical hurdles to implementing such a task can be easily overcome. Software will need to be devised to funnel information from each type of analysis to each respective database and be put into useful ordered text file locations. These files would contain information detailing location of the gene on a particular chromosome, number of exons, splice sites, transcriptional boundaries, promoter sites, and other control regions. To verify such an approach I would choose to use the yeast genome as a starting point to train the entire process. The genome is fairly small compared to humans, has been well documented, and functional studies of the entire ORFeome are underway. This could provide a basis for comparing the output from such a study with those previously attempted. Given that the yeast genome contains very few multi-exon genes the annotation scheme must be tested on additional genomes to streamline and enhance the process. Using portions of the human genome that have already been well documented

could provide such a test set. Once the scheme had proved its worth it could then be unleashed onto the entire genome without restriction to create a global, genome wide annotation. Such an integrated approach has not been attempted thus far. Previous annotations were very limited in their scope and thereby limited in the types of genetic information they could uncover. By using all of the structural and functional genetic knowledge we have to day such an integrated approach could give a working annotation of the human genome that is far more complete than anything presently available.

1. J. Aach *et al.*, *Nature* **409**, 856 (Feb 15, 2001).
2. L. Stein, *Nat Rev Genet* **2**, 493 (Jul, 2001).
3. J. Wang *et al.*, *Nat Rev Genet* **4**, 741 (Sep, 2003).
4. M. Q. Zhang, *Nat Rev Genet* **3**, 698 (Sep, 2002).
5. N. L. Harris, *Mol Biotechnol* **16**, 221 (Nov, 2000).
6. S. Oliver, *Trends Genet* **12**, 241 (Jul, 1996).
7. M. Johnston, *Trends Genet* **12**, 242 (Jul, 1996).
8. A. Goffeau *et al.*, *Science* **274**, 546 (Oct 25, 1996).
9. B. Dujon, *Trends Genet* **12**, 263 (Jul, 1996).
10. A. Bateman *et al.*, *Nucleic Acids Res* **32**, D138 (Jan 1, 2004).
11. B. Boeckmann *et al.*, *Nucleic Acids Res* **31**, 365 (Jan 1, 2003).
12. C. J. Sigrist *et al.*, *Brief Bioinform* **3**, 265 (Sep, 2002).
13. H. M. Berman *et al.*, *Nucleic Acids Res* **28**, 235 (Jan 1, 2000).
14. V. Matys *et al.*, *Nucleic Acids Res* **31**, 374 (Jan 1, 2003).
15. C. Amoreira, W. Hindermann, C. Grunau, *Nucleic Acids Res* **31**, 75 (Jan 1, 2003).