Gene finding using multiple related species: a classification approach.

Manolis Kellis

Special review for the Encyclopedia of Genomics, Proteomics, and Bioinformatics.
John Wiley and Sons, editors. (in press)

Three years after the initial sequencing of the human genome, the actual number of functional human genes remains uncertain. Several expression-based analyses still argue for a hundred thousand transcribed genes, whereas more conservative estimates range between 20 and 25 thousand genes. The central question in such debates still remains: what constitutes a *real* gene? In this paper, we address this question and present a comparative genomics approach for systematic gene identification, which observes gene-specific signatures of evolutionary selection across multiple related species. First, we formulate the gene identification problem as a classification problem between genes and non-coding regions, based on their distinct patterns of nucleotide change. We then summarize the results of applying this approach to re-annotate the yeast genome, with changes affecting nearly 15% of all genes, and the rejection of more than 500 previously annotated genes. Finally, we discuss the implications of this analysis on understanding the human genome, and strategies for the systematic reannotation of higher eukaryotes.

Ideally, we should be able to systematically discover all the functional genes in a newly sequenced genome, from its sequence alone. Computational discovery methods rely both on the direct signals used by the cell to guide transcription, splicing, and translation, and also on indirect signals such as evolutionary conservation. In this paper, we summarize the principles of a classification-based approach for systematic gene identification, based on comparative sequence information from multiple closely related species. We first frame gene identification as a classification problem of distinguishing real genes from spurious gene predictions. We then present the Reading Frame Conservation test[1], a new computational method implementing such a classification approach, based on the patterns of nucleotide changes in the alignment of orthologous regions. We finally summarize our results of applying this method to re-annotate the yeast genome[2], and the challenges of using related methods to discover all functional genes in the human genome.

Defining real genes

What is a real gene? This question is relatively easy to answer for those genes that are abundantly expressed, encode well-characterized proteins, and whose disruption affects a specific function in the organism. Beyond these, the distinction is much more subtle between functional genes and spurious gene predictions.

Experimentally, the definition of a functional gene comes largely as accumulating evidence of its usage, including gene expression[3], protein fragments[4], biochemical function[5], protein interactions[6], or the effect of its disruption[7]. It is important to note, however, that absence of experimental evidence does not imply that a gene annotation is spurious: a real gene may be missing experimental evidence because it is not used in the particular conditions surveyed. Conversely, any individual report of gene usage could be due to experimental noise, cross-hybridization, or chance transcription events due to a basal level of intergenic transcription. Thus, experimental evidence alone is insufficient to distinguish between real genes and spurious gene predictions.

Computationally, the processes of transcription, splicing and translation can be thought of as a series of decisions taken by the cell, based on signals in the genome. These signals include distant enhancer elements, regulatory elements surrounding the transcription start site, splicing enhancer and repressor signals in the message, and translation signals in the sequence or structure of mature mRNAs[8-11]. The subset of these signals that we currently understand is insufficient to specify the set of known genes. Thus, in addition to the *direct* signals used by the cell, gene identification methods[12-18] routinely identify genes using additional *indirect* signals (reviewed in [19]), which are not available to the cell, although they are generally good indicators of protein-coding genes. These include the frequency of each codon in protein-coding regions, the overall length of the translated protein product, and importantly the evolutionary conservation of protein sequences across related species.

Evolutionary conservation is perhaps the strongest indicator that a predicted gene is functional. A gene that confers even a slight evolutionary advantage can be conserved across millions of years, regardless of the rarity of its usage. Hence, even if experimental methods fail to detect the usage of a gene in a given set of experimental conditions, evolutionary methods are able to detect *indirect* evidence of its usage, by detecting pressure to preserve its function over millions of years.

Gene finding as a classification problem

The challenge of using evolutionary conservation to define genes lies in the ability to reject spurious genes. Although it is generally well understood that genes conserved across large evolutionary distances are functional, lack of conservation is generally attributed to evolutionary divergence, thus lacking evidence towards either accepting or rejecting a gene prediction.

Yet, the comparison of related genomes contains information much richer than simply the presence or absence of a protein in a genome. By working with closely related genomes, we can define regions of conserved gene order, or *synteny blocks*, which span several genes, and potentially entire chromosomes. Defining conserved synteny blocks allows us to construct global alignments, spanning both well-conserved regions, but also regions of low conservation. In particular, these give us access to full nucleotide alignments for all predicted genes, and for all intergenic regions within orthologous segments.

With the availability of orthologous alignments for both protein-coding and non-coding regions, we can study their distinct properties and build a classifier between the two types of region. The simplest and most commonly used such classifier observes the overall level of nucleotide conservation in the alignment, and selects regions of high conservation as likely to be functional. Other classifiers may observe more subtle signals, such as the number of amino acid changes per nucleotide substitution[20], the frequency of insertions and deletions, the periodicity of mutations, and so on. By working with properties of the nucleotide alignment, rather than protein alignments, we are able to apply it uniformly to evaluate both genes and intergenic regions in the same test.

One particular conservation property unique to protein-coding segments is the pressure to preserve the reading frame of translation. Since protein sequences are translated every three nucleotides, the length of insertions and deletions (indels) is largely constrained to remain a multiple of three, thus preserving the frame of translation. Within coding regions, indels which disrupt the frame of translation are excluded, or compensated with nearby indels which restore

the reading frame. In non-coding regions, the length of indels does not have this constraint, and short spacing changes are tolerated.

To evaluate this property quantitatively, we developed the Reading Frame Conservation (RFC) test. This test evaluates the pressure to preserve reading frame in a fully aligned interval, by measuring the portion of nucleotides in this interval for which the reading-frame has been locally conserved[1]. The RFC test provides a classifier between coding and non-coding regions that is completely independent of the traditional signals used in gene identification. It does not rely on start, stop, or splicing signals, nor does it rely on the conservation of protein sequence. It can therefore be combined with existing gene finding tools and provide a highly informative score for any interval considered.

Reannotation of yeast

A classification approach is particularly well-suited for the yeast genome. The general scarcity of introns makes it hard to rely on splicing signals to discover genes. Thus, the annotation of S. cerevisiae has traditionally relied solely on the length of predicted proteins to annotate genes, resulting in 6062 annotated open reading frames (ORFs) which potentially encode proteins of at least 100 amino acids (aa). Additionally, a tentative functional annotation has been inferred for as many as 3966 of these ORFs, based on classical genetic experiments and systematic genome-wide studies of gene expression, deletion phenotype, and protein-protein interaction. Together, the interval-based annotation, and the large set of well-known genes make it possible to apply the RFC test systematically to evaluate the functional significance of the remaining ORFs.

To apply the test, we constructed multiple sequence alignments for every ORF and every intergenic region, across four closely related species. We sequenced and assembled the genomes of *S. paradoxus, S. mikatae, and S. bayanus*[2], and defined genome-wide synteny blocks with *S. cerevisiae*, based on discrete anchors provided by unique protein blast hits[1]. Within these synteny blocks, we constructed global alignments of orthologous genes and intergenic regions across the four species using CLUSTALW[21], and systematically evaluated each alignment. We compared *S. cerevisiae* to each species in turn, and every comparison cast a vote based on its overall RFC and a species-specific cutoff. A decision was then reached for each gene by tallying the votes from all comparison.

We evaluated the sensitivity and specificity of the approach based on the 3966 genes with functional annotations ('known genes'), and 340 randomly chosen intergenic sequences with lengths similar to the annotated ORFs. The RFC test correctly accepted 3951 known genes (99.6%) and rejected only 15 known genes; upon manual inspection, these 15 are indeed likely to be spurious (most lack experimental evidence, and deletion phenotypes of the rest are likely due to their overlap with the promoter of other known genes). The method also correctly rejected 326 intergenic regions (96%), accepting only 14 intergenic regions (of which 10 appear to define short ORFs or extend annotated ORFs, suggesting that at most 1% of true intergenic regions failed to be rejected by the RFC test). In summary, the RFC test shows a very strong discrimination between genes and intergenic regions, with sensitivity and specificity values greater than 99%.

We then applied the RFC test to all previously annotated genes, leading to a major revisiting of the yeast gene set. For ORFs with lengths greater than 100 aa, our analysis accepted 5538 ORFs and rejected 503 (of which 376 were immediately rejected, 105 were rejected with additional criteria, and 32 were merged with neighboring ORFs); the classifier abstained from making a

decision in 20 cases. The rejected ORFs show an abundance of frame-shifting indels across their entire length, in-frame stop codons, and low conservation of protein sequence, in addition to the low RFC score; their length distribution and atypical codon usage additionally suggest that they are likely occurring by chance[22-24]; furthermore previous systematic experimentation showed no compelling evidence that these may encode a functional gene. Thus, it appears that more than 500 previously annotated ORFs in the yeast genome are spurious predictions.

Below the 100 amino acid cutoff, no previous systematic annotation or experimentation was available. Thus, to validate our method, we compared the results of the RFC test with an independent metric: the proportion of the *S. cerevisiae* ORF that was also free of stop codons in the other three species. Between 50 and 99 aa, we found that the method is still reliable, and report 43 candidate new genes at that length. As ORFs smaller than 50 aa were tested, we found that the specificity of the test decreased, since small intervals tend to be devoid of indels by chance rather than the presence of selective pressure. Thus, additional constraints, and additional species, will be needed to discover genes reliably at such short lengths.

In addition to the discovery of genes themselves, the comparative analysis allowed us to refine gene boundaries. Once an interval was determined to be under selective pressure for reading frame conservation, the boundaries of that interval were adjusted, based on the conservation of start/stop/splice signals, and the boundaries over which reading frame is conserved. This led to a large-scale reannotation of gene structure, which affected hundreds of genes (146 start codon changes, 67 intron changes, 32 merges of consecutive ORFs, and 45 changes of ORF ends). It is worth noting that in 134 cases, the inferred boundary changes pinpointed sequencing errors in the primary sequence of *S. cerevisiae*, many of which were tested and corrected by re-sequencing. These boundary changes reveal the true location of promoter regions, new protein domains in elongated ORFs, and previously overlooked functional relationships in the case of merged ORFs

In summary, the RFC test was able to reliably distinguish between genes and intergenic regions, leading to a systematic reannotation of the yeast genome. The comparative analysis led to the rejection of more than 500 previously annotated genes and to the discovery of many novel genes. The results agree with similar comparative analyses carried out from a number of yeast species[25-28]. In addition, it allowed us to refine the gene structure of hundreds of genes, adjusting start and stop boundaries, merging consecutive ORFs, and discovering many new introns. Moreover, by using multiple species, the signals leading to gene identification were powerful enough to suggest sequence changes in any individual species, even in *S. cerevisiae*; we tested ~50 of these predicted sequencing errors in *S. cerevisiae*, confirming and correcting the large majority of them by resequencing.

Implications for the human genome

The challenge of gene finding in the human genome is far greater than in yeast, due to the vastly larger intergenic regions, numerous and large introns, small exons, and alternatively spliced genes. Most approaches to gene finding in higher eukaryotes have relied on Hidden Markov Models[12-17], which inherently emphasize the importance of exon chaining, and rely on knowledge of the expected length distributions for both exons and introns. These have been recently extended to use sequence from multiple species[29-38]. These approaches have limitations however, in cases of alternative splicing[39-41], differences in splicing between species[42], widely varying exon and intron lengths, and non-canonical splice sites.

An exon-based classification approach to gene finding can help overcome these limitations. By exhaustively enumerating and testing all candidate protein-coding intervals, classification approaches make exon detection independent of chaining. Splice site models can be applied to each species independently and then compared to each other, which is much stronger than the evidence in any one species alone. The intervals can be then tested based on discriminating variables that distinguish genes from non-coding regions[19,43]. New discriminating variables can be defined as alignments of each type of region are systematically compared, and these can be combined and weighted, leveraging traditional machine learning techniques for feature selection and classification[44]. Once relevant intervals have been identified, their boundaries can be adjusted for optimal chaining into complete genes. In particular, chaining can also leverage an inferred frame of translation for each exon, based on the higher mutation rate observed in largely degenerate third codon positions[20]. The exon-chaining step produces full gene models, and is able to cope with alternative splicing and missing exons, since it is not constrained to a single optimal path through the exons.

By systematically observing alignment properties of large sequence regions, we can build new rigorous approaches for sequence analysis. These are widely applicable beyond coding exons, and similar classification-based approaches can be used to distinguish CpG islands, 5'- and 3'-untranslated regions, promoter regions, regulatory islands, and other functional elements. Through the lens of evolutionary selection, our ability to directly interpret genomes is revolutionized. Coupled with systematic experimentation and validation, these analyses can lead to a systematic catalogue of functional elements in the human genome, forming the future foundations of biomedical research.

References

1. Kellis, M., Patterson, N., Birren, B., Berger, B. & Lander, E. S. Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *J Comput Biol* **11**, 319-55 (2004).
2. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241-54 (2003).
3. Velculescu, V. E. et al. Characterization of the yeast transcriptome. *Cell* **88**, 243-51 (1997).
4. Rezaul, K., Wu, L., Mayya, V., Hwang, S. I. & Han, D. K. A systematic characterization of mitochondrial proteome from a human T leukemia cells. *Mol Cell Proteomics* (2004).
5. Jackman, J. E., Montange, R. K., Malik, H. S. & Phizicky, E. M. Identification of the yeast gene encoding the tRNA m1G methyltransferase responsible for modification at position 9. *Rna* **9**, 574-85 (2003).
6. Bai, C. & Elledge, S. J. Gene identification using the yeast two-hybrid system. *Methods Enzymol* **283**, 141-56 (1997).
7. McAlister, L. & Holland, M. J. Targeted deletion of a yeast enolase structural gene. Identification and isolation of yeast enolase isozymes. *J Biol Chem* **257**, 7181-8 (1982).
8. Fairbrother, W. G., Yeh, R. F., Sharp, P. A. & Burge, C. B. Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007-13 (2002).
9. Wang, Z. et al. Systematic identification and analysis of exonic splicing silencers. *Cell* **119**, 831-45 (2004).
10. Thanaraj, T. A. & Robinson, A. J. Prediction of exact boundaries of exons. *Brief Bioinform* **1**, 343-56 (2000).

11. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**, 377-94 (2004).
12. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78-94 (1997).
13. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878-9 (2004).
14. Kulp, D., Haussler, D., Reese, M. G. & Eeckman, F. H. A generalized hidden Markov model for the recognition of human genes in DNA. *Proc Int Conf Intell Syst Mol Biol* **4**, 134-42 (1996).
15. Krogh, A. Two methods for improving performance of an HMM and their application for gene finding. *Proc Int Conf Intell Syst Mol Biol* **5**, 179-86 (1997).
16. Henderson, J., Salzberg, S. & Fasman, K. H. Finding genes in DNA with a Hidden Markov Model. *J Comput Biol* **4**, 127-41 (1997).
17. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**, II215-II225 (2003).
18. Salzberg, S., Delcher, A. L., Fasman, K. H. & Henderson, J. A decision tree system for finding genes in DNA. *J Comput Biol* **5**, 667-80 (1998).
19. Fickett, J. W. & Tung, C. S. Assessment of protein coding measures. *Nucleic Acids Res* **20**, 6441-50 (1992).
20. Hurst, L. D. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet* **18**, 486 (2002).
21. Higgins, D. G. & Sharp, P. M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237-44 (1988).
22. Goffeau, A. et al. Life with 6000 genes. *Science* **274**, 546, 563-7 (1996).
23. Dujon, B. et al. Complete DNA sequence of yeast chromosome XI. *Nature* **369**, 371-8 (1994).
24. Sharp, P. M. & Li, W. H. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**, 1281-95 (1987).
25. Cliften, P. et al. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* **301**, 71-6 (2003).
26. Blandin, G. et al. Genomic exploration of the hemiascomycetous yeasts: 4. The genome of Saccharomyces cerevisiae revisited. *FEBS Lett* **487**, 31-6 (2000).
27. Wood, V., Rutherford, K. M., Ivens, A., Rajandream, M.-A. & Barrell, B. A Re-annotation of the *Saccaromyces cerevisiae* Genome. *Comparative and Functional Genomics* **2**, 143-154 (2001).
28. Brachat, S. et al. Reinvestigation of the Saccharomyces cerevisiae genome annotation by comparison to the genome of a related fungus: Ashbya gossypii. *Genome Biol* **4**, R45 (2003).
29. Yeh, R. F., Lim, L. P. & Burge, C. B. Computational inference of homologous gene structures in the human genome. *Genome Res* **11**, 803-16 (2001).
30. Siepel, A. & Haussler, D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol* **21**, 468-88 (2004).
31. Siepel, A. & Haussler, D. Combining phylogenetic and hidden Markov models in biosequence analysis. *J Comput Biol* **11**, 413-28 (2004).
32. Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B. & Lander, E. S. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res* **10**, 950-8 (2000).
33. Korf, I., Flicek, P., Duan, D. & Brent, M. R. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 Suppl 1**, S140-8 (2001).

34. Dewey, C. et al. Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat. *Genome Res* **14**, 661-4 (2004).
35. Rinner, O. & Morgenstern, B. AGenDA: gene prediction by comparative sequence analysis. *In Silico Biol* **2**, 195-205 (2002).
36. Parra, G. et al. Comparative gene prediction in human and mouse. *Genome Res* **13**, 108-17 (2003).
37. Meyer, I. M. & Durbin, R. Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics* **18**, 1309-18 (2002).
38. Novichkov, P. S., Gelfand, M. S. & Mironov, A. A. Gene recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics* **17**, 1011-8 (2001).
39. Brett, D. et al. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett* **474**, 83-6 (2000).
40. Mironov, A. A., Fickett, J. W. & Gelfand, M. S. Frequent alternative splicing of human genes. *Genome Res* **9**, 1288-93 (1999).
41. Cawley, S. L. & Pachter, L. HMM sampling and applications to gene finding and alternative splicing. *Bioinformatics* **19 Suppl 2**, II36-II41 (2003).
42. Nurtdinov, R. N., Artamonova, II, Mironov, A. A. & Gelfand, M. S. Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum Mol Genet* **12**, 1313-20 (2003).
43. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11**, 725-36 (1994).
44. Moore, J. E. & Lake, J. A. Gene structure prediction in syntenic DNA segments. *Nucleic Acids Res* **31**, 7271-9 (2003).