Abstracts

Briefings in Bioinformatics aims to provide working biologists with an awareness and understanding of the computational approaches available for research and discovery. The Abstracts section of the journal consists of summaries of bioinformatics manuscripts published in the previous quarter. Inclusion of an article in this section indicates that the editors consider it to be among the most interesting and/or useful contributions to the field for the quarter covered. The contents of these reports are briefly distilled for the readers with an emphasis placed on their biological context and potential utility. Publications from the first quarter of 2006 (January–March) are reviewed here.

Genome-wide identification of human functional DNA using a neutral indel model

Gerton Lunter, Chris P. Ponting and Jotun Hein

PLoS Computational Biology (2006) Vol. 2, no. 1, p. e5

The vast majority of human genomic sequence is not protein coding, and putative functional roles for almost all of this sequence have yet to be discovered. In light of this knowledge gap, there is a sustained effort underway to develop computational approaches to aid in functional predictions for non-coding DNA. In the last several years, the importance of using comparative (between genome) sequence analysis in such efforts has become increasingly obvious. This is because functionally important sequences tend to evolve more slowly than non-functional sequences and thus can be identified by virtue of their low levels of sequence divergence. Lunter *et al*. propose a corollary to this approach by focusing on insertion and deletion (indel) events as opposed to nucleotide substitutions. The first part of their approach involves the derivation of the neutral background expectation for levels of indel when no selection is acting on the sequences. The model they devise towards this end shows a good fit to the indel pattern for human–mouse ancestral repeats, which are not thought to evolve under selection. Having defined the neutral indel rate in this way, the authors were then able to identify a number of long ungapped regions (significantly non-neutral) that they presume to be greatly enriched for functionally active non-coding DNA. Fully 2.5–3.2% of the human genome – about 2 times as much as coding regions – appears to have evolved under a regime of purifying selection against indel events. While many of the sequences in these selected regions are not annotated, the method performs quite well at identifying know functional elements such as microRNAs. One of the substantial benefits of their method is that it seems to require much lower genome coverage than comparative methods that evaluate sequence substitutions. Comparison of only 2–3 mammalian genomes is sufficient to identify 90 million bases of human DNA subject to purifying indel selection with predicted levels of only 10% false positives and 25% false negatives. Finally, the authors point out that application of their method can lead to the detection of genomic regions that have been subject to heterogeneous selection, like purifying selection against indels with diversifying selection on sequence substitutions. This raises the possibility of extending genome-scale surveys of positive selection from protein coding regions to non-coding DNA, an important development in molecular evolution.

Stable evolutionary signal in a Yeast protein interaction network

Stefan Wuchty, Albert-Laszlo Barabási and Michael T. Ferdi

BMC Evolutionary Biology (2006)
Vol. 6, no.1, p. 8

Network based analyses have helped to provide an integrated systems-level perspective on the functional and evolutionary relationships among thousands of genes. One of the first examples of this unified approach was the discovery that highly connected proteins, so-called hubs of protein interaction networks, tend to be more evolutionarily conserved than less prolific interactors. However, this seemingly straightforward claim has been controversial and subject to great scrutiny in the past few years. Much of the ensuing debate has focused on problems with the reliability of high-throughput protein interaction data sets and the accuracy of ortholog assignments based on sequence comparison. Wuchty and colleagues confirm that the evolutionary conservation of highly connected proteins is robust to both of these potential sources of noise in large-scale genomic data sets. Perhaps more importantly, their work entails an important qualitative shift in focus away from a node-centric perspective to a more link centered one. In other words, the emphasis here is placed on the conservation of interactions between gene products as opposed to the conservation of the genes (proteins) themselves. In so doing, the authors are able to discover that local network substructure, in terms of clustering around interactions, is correlated with evolutionary conservation of the proteins involved in the interactions. In addition, the genes found in these locally clustered sub-networks show elevated levels of coexpression consistent with *bona fide* functional relationships among them. One of the most intriguing things to come out the work is an expansion of the notion of the fundamental evolutionary unit from single proteins to ensembles proteins and the interactions among them. This has theoretical implications for the understanding of the action of natural selection as well as potential practical utility with respect to the prediction of interaction networks (*i.e.* functional information transfer) for non-model organisms. Indeed, the authors use a comparison of local interaction network structure and orthology between *Saccharomyces cerevisiae* and *Plasmodium falciparum* to demonstrate the potential of such predictive methods.

Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana

Tineke Casneuf, Stefanie De Bodt, Jeroen Raes, Steven Maere and Yves Van de Peer

Genome Biology (2006) Vol. 7, no. 2, p. R13

Gene duplication is an important source of genetic novelty and has been particularly prevalent in plants such as the model system *Arabidopsis thaliana*. Comparative studies of the Arabidopsis genome sequence have revealed that it has experienced several rounds of large-scale, whole genome duplication in addition to many smaller, local duplication events. It has previously been shown that Arabidopsis genes often acquire distinct expression patterns after duplication but the roles of gene function and duplication age in this process have not been well appreciated. Casneuf *et al*. study the effects of different classes of duplication on patterns of gene expression and consider the relationship between expression divergence and the encoded functions of the duplicated genes. All-against-all sequence comparisons were used to detect duplicates, and the relative ages of gene duplications were defined by comparing the synonymous substitution divergence between duplicated genes. The authors then analyzed two publicly available Arabidopsis microarray data sets to assess the extent of expression pattern divergence between duplicate genes. They found that genes that duplicated via local small scale events had acquired relatively divergent patterns of expression compared to those duplicated in large scale events. Locally duplicated gene pairs also tended to diverge asymmetrically in the sense that one duplicate was expressed in a large number of tissues while the other was expressed in fewer tissues. It is not clear whether this asymmetric divergence pattern is due to a distinct sub-functionalization of the duplicate gene with the more narrow expression pattern or whether that same gene is in the process of losing much of its function (becoming a pseudogene). The comparison of pairs of genes also limits the interpretation of the results in the sense that the authors could not determine if the asymmetric expression patterns were due to gain or loss of tissue-specific expression. Nevertheless, the difference between the two modes of expression divergence for the two classes of duplicates is quite striking. In terms of function, the expression patterns of genes involved in signal transduction and response to external stimuli diverge relatively quickly. This could be an evolutionary mechanism by which *Arabidopsis* has evolved to meet the challenges of a changing environment.

Transposon-free regions in mammalian genomes

Cas Simons, Michael Pheasant, Igor V. Makunin and John S. Mattick

Genome Research (2006) Vol. 16, no. 2, pp. 164–172

Mammalian genomes are made up primarily of transposable elements (TEs) – mobile sequences capable of replicating themselves independent of their host genomes. The human and mouse genomes, for instance, consist of more than 50% TE-derived sequence, made up of millions of individual elements that are separated by only 500bp on average. Despite the ubiquity and abundance of mammalian TEs, there are regions of the genome that are entirely devoid of these elements. Simon *et al.* focus on these so-called 'transposon-free regions' (TFRs), and their analysis of these anomalous genomic regions has important implications for understanding mechanisms of gene regulation. They show that the human and mouse genomes both have close to 1,000 TFRs more than 10kb in length, and most of the TFRs are present as corresponding orthologous regions in the human and mouse genomes. The latter fact is particularly interesting since most mammalian TEs are lineage-specific, in other words they have accumulated since the human and mouse genomes last shared a common ancestor. Thus, the presence of orthologous TFRs means that TE sequences have been selectively excluded – *i.e.* prevented from inserting – independently in human and mouse orthologous regions. This suggests similar selective constraint for orthologous human–mouse TFR regions, which must of course be based on some functional role for those genomic sequences. More than 90% of the TFR sequences are non-coding so the function of those selected regions is probably related to gene regulation. Consistent with this interpretation, most TFRs are found in the vicinity of genes involved in developmental regulation. Paradoxically, however, the orthologous TFRs are not highly conserved in terms of sequence. This is most surprising because it is expected that orthologous, functionally analogous regions will be anomalously conserved in terms of the level of sequence divergence. Indeed, this expectation is the basis of a widely employed class of approaches, known as phylogenetic footprinting, which use sequence conservation levels to guide the identification of non-coding functional genomic elements. This work, on the other hand, suggests a novel view of regulatory sequence evolution whereby genomic regions are somehow selected for regulatory function without conservation of their particular genome sequence.

Evolution of alternative splicing after gene duplication

Zhixi Su, Jianmin Wang, Jun Yu, Xiaoqiu Huang and Xun Gu

Genome Research (2006) Vol. 16, no. 2, pp. 182–189

The importance of gene duplication for the evolution of novel biochemical functions has long been recognized. The basic model for this process holds that duplication provides the redundancy needed to free single copy genes from selective constraint, and most studies to date have focused on how this redundancy facilitates changes in sequences (*i.e.* point substitutions) between duplicate gene copies. Here, Su and colleagues investigate the role of alternative splicing (AS) in this process by evaluating AS differences between duplicate human genes. Previous studies found that gene duplication reduces AS suggesting that these two phenomena may represent complementary modes for generating diversity. However, there is much that remains to be learned about the relationship between gene duplication and AS. The current study addresses a number of outstanding questions including: i-how long does it take for AS to evolve after duplication, ii-is AS of duplicates asymmetric and iii-does AS evolve independently of gene duplication. Consistent with previous results, the authors did find that duplicate genes have fewer AS variants than single copy genes and members of larger gene families tend to have even fewer AS variants. It was also demonstrated that the loss of AS variants can occur shortly after gene duplication. Furthermore, there is pronounced asymmetry in the patterns of AS between duplicate genes whereby duplicates have pronounced differences in their splicing patterns. All of this points to a tightly coupled and complementary relationship between gene duplication and AS. Apparently, in the period immediately following gene duplication the duplicates take on distinct aspects of the function that was previously carried out by the single copy ancestral gene via AS. This type of reductive evolution is known as subfunctionalization. It is only later on that the evolution of AS between duplicate gene copies occurs more-or-less independently.

A genome-wide study of dual coding regions in human alternatively spliced genes

Han Liang and Laura F. Landweber

Genome Research (2006) Vol. 16, no. 2, pp. 190–196

The prevalence and importance of alternative splicing (AS) of eukaryotic genes is by now quite well appreciated. In this paper, Liang and Landweber analyze AS genes that encode multiple protein sequences from overlapping reading frames. This phenomenon is far less recognized among eukaryotes, and was previously considered to be largely confined to small prokaryotic and viral genomes. For such small genomes, overlapping reading frames represent an evolutionary strategy by which genomes can increase their coding content without concomitant increases in size. Since there is far less apparent selective pressure to streamline eukaryotic genomes, overlapping reading frames were presumed to be less prevalent. By closely examining a set of human mRNA transcript sequences, the authors were able to show that just under 7% of alternatively spliced genes contain multiple overlapping coding frames. It is important to note that the authors took measures to ensure the reliability of the analyzed transcript set, focusing only on relatively well-annotated sequences. The figure of 7% is slightly mitigated in that it represents only ~1% of all human genes analyzed, nevertheless this still seems to be a surprisingly high figure for such a large eukaryotic genome and suggests the emergence of multiple overlapping reading frames is an important evolutionary force in eukaryotes. Interestingly, many of these secondary reading frames appear to have emerged recently, being mammalian and even primate-specific. The derived reading frames have significantly different codon usage and amino acid composition than ancestral frames and the rest of the genes in the genome. At first glance, these latter findings may seem to suggest that the derived secondary reading frames encode non-functional proteins, and this is a critical point regarding the significance of the work. However, the authors performed a number of comparative sequence analyses that suggest that most of these transcripts encode proteins that are at least potentially functional. Specifically, the authors tested whether proteins encoded in secondary reading frames were similar to known functional sequence from other organisms, whether they contained functional motifs, whether they could form know secondary structural elements and whether they could be found to be similar to distantly related structures via threading. Almost all proteins encoded by overlapping reading frames conform to at least one of these criteria and most meet multiple lines of evidence for function.

"Genome design" model: Evidence from conserved intronic sequence in human–mouse comparison

Alexander E. Vinogradov

Genome Research (2006) Vol. 16, no. 3, pp. 347–354

Eukaryotic genome sequences are saddled with vast regions that do not encode any known protein; for instance, the human genome is 98–99% non-coding. One of the fundamental challenges of the post-genomic era is to characterize the functional significance, or lack thereof, for this non-coding DNA in eukaryotic genomes. One prominent class of non-coding DNA is introns, and several recent studies have show that intron lengths vary substantially between broadly expressed house-keeping genes, which possess relatively shorter introns, and genes with longer introns that are expressed in more tissue-specific and/or developmentally-specific patterns. There are two distinct classes of evolutionary explanations, one neutral and one adaptationist, for the differences in intron lengths between these two groups of genes. The neutral explanation holds that selection acts on the more highly expressed house-keeping genes to keep them small and streamlined while the less expressed tissue-specific genes are more free to accumulate intronic DNA. At this moment, the neutral, or 'selection for economy', model corresponds to more-or-less to received wisdom. Vinogradov, on the other hand, finds support for the less invoked adaptationist perspective in his comparative analysis of mammalian intron and expression data. He refers to this as the genome design hypothesis, which suggests that tissue-specific genes possess more intronic DNA due to its functional role in their complex regulation. What he did was employ some of the most reliable technology in local sequence alignment to compare the amount of intron length conserved between orthologous human and mouse genes and considered these data with respect to the relative breadth of gene expression. Use of these rigorous sequence alignment algorithms allowed for the discovery of the highest fractions of observed human–mouse conserved intronic DNA, excluding lineage-specific repeats, to date. More to the point, Vinogradov shows that the fraction of conserved intronic sequences is higher for tissue-specific genes than for house-keeping genes, and this observation holds up after a number of controls are implemented, e.g. for mutation rate and GC content. Intron length was also found to be correlated with the number of functional domains in a gene. Another interesting finding was that the length distributions for conserved intron sequences peak close to the nucleosomal (170–220nt) and dinucleosomal (320–510nt) intervals. Taken together, these results suggest a possible connection between intron sequences and gene regulation mediated by chromatin condensation.

Toward automatic reconstruction of a highly resolved tree of life

Francesca D. Ciccarelli, Tobias Doerks, Christian von Mering, Christopher J. Creevey, Berend Snel and Peer Bork

Science (2006) Vol. 311, no. 5765, pp. 1283–1287

Ever since Darwin put forth the notion that all of life has descended with modification from a common ancestral origin, the reconstruction of a 'tree-of-life', which defines the evolutionary relationships among species, has been a fundamental goal for the biological sciences. The molecular revolution showed that all species, no matter how phenotypically distinct, were unified at the genomic level and made the realization of this goal a formal possibility. In recent years, as numerous whole genome sequences have become available, a number of investigators have attempted, with notable success, to reconstruct a comprehensive tree-of-life. Despite these recent advances however, there remain fundamental technical and conceptual challenges to this endeavor. For instance, the precise phylogenetic placement of numerous taxonomic groups remains hotly debated, and even the basic notion that species evolution can be accurately represented by a tree-like structure has been contested. Ciccarelli *et al*. address this issue by devising an automated procedure for building a tree-of-life that takes advantage of complete genome sequences and, critically, is reproducible and readily updated. They start with close to 200 genome sequences, and the first task is to identify conserved gene families present in all genomes and establish 1:1 orthology among members of these families. The resulting set consists of 31 universal genes all of which are involved in translation (ribosomal proteins for the most part). It should be noted that the identification and extraction of this universally conserved gene set required a substantial amount of manual analysis; however, having set up the procedure in this way, subsequent work can be completed in a largely automated fashion. An important part of the analysis and computational pipeline involves the identification and removal of genes that have evolved via horizontal transfer – such genes will inevitably lead to inconsistencies in any attempt to resolve a tree of life. Preliminary analysis led to the identification, and removal, of five families that were universally present in all 191 species but showed evidence of multiple horizontal transfer. Many of these genes were already known to have evolved by horizontal transfer. Less expected was the finding that a number of the universally conserved ribosomal genes showed evidence of horizontal transfer. After controlling for these events, the result was a concatenated alignment of more than 8,000 positions and this alignment was used to reconstruct a universal tree of life. The utility of their approach is underscored by the fact that the internal branches on the resulting tree of life have remarkably good statistical (bootstrap) support. This robust branching allowed the authors to propose resolutions to several outstanding taxonomic issues. For instance, they find evidence for a Gram-positive origin of bacteria and for a thermophilic last universal common ancestor.

A third approach to gene prediction suggests thousands of additional human transcribed regions

Gustavo Glusman, Shizhen Qin, M. Raafat El-Gewely, Andrew F. Siegel, Jared C. Roach, Leroy Hood and Arian F. A. Smit

PLoS Computational Biology (2006)
Vol. 2, no. 3, p. e18

Only a small fraction (1–2%) of the human genome encodes protein sequences, but there is evidence that much, if not the majority, of the genome is actually transcribed. Computational discovery of functionally relevant transcribed regions is an important goal of functional genomics. Sustained transcription of genomic sequences has pronounced effects such as the creation of differences between the two strands of genomic DNA. Specifically, there are mutational biases caused by transcription coupled DNA repair as well as stand-specific selection against polyadenylation signals. The authors of this report propose to take advantage of these so-called 'transcription footprints' to identify previously undetected genes that do not encode protein sequences. This approach represents a third and qualitatively distinct method of gene prediction – the most widely used methods rely on detection of gene structural elements and sequence conservation. As such, their method is complementary to existing gene prediction algorithms and it has the potential to reveal many previously undiscovered genes in mammalian genomes. The specific approach described here involves the integration of four specific algorithms, each of which detects some aspect of mutational and/or strand bias caused by transcription. Importantly, the signals that these algorithms pick up are confined to sequences that have undergone sustained transcription over evolutionary time as opposed to noise cause by spurious and ephemeral transcription. Application of this integrated algorithm resulted in the detection of thousands of transcribed genomic segments that do not correspond to currently known genes. Depending on the statistical threshold employed, their method indicates that as much as half of the human genome is consistently transcribed. While their method performs generally well when tested on known gene sequences, it appears to be biased towards the detection of long transcribed regions. This may allow their approach to help in the detection of genes with anomalously long introns that are resistant to discovery with the most commonly used methods based on gene structure and sequence similarity.

Gene losses during human origins

Xiaoxia Wang, Wendy E. Grus and Jianzhi Zhang

PLoS Biology (2006) Vol. 4, no. 3, p. e52

Historically, the emphasis in understanding the increases in complexity that accompanied the emergence of the human evolutionary lineage has been placed on the creation of novel genetic and biochemical function. However, a recent hypothesis regarding evolutionary innovation – referred to as "less-is-more" – turns this paradigm upside down by implicating gene loss as an important force in evolutionary innovation. In support of this notion, several anecdotal studies have pointed to specific cases where gene loss (pseudogenization) may have led to adaptations related to human-specific characteristics. To date, however, there has been no systematic attempt to characterize gene loss along the human lineage along with the possible evolutionary implications of this process. Wang *et al.* have conducted just such a systematic analysis by attempting to characterize the full repertoire of human-specific pseudogenes. This was accomplished by comparative human-chimp genome sequence analysis together with an in depth literature analysis. Their study focused on nonprocessed pseudogenes, rather than processed pseudogenes, which are more abundant but less functionally relevant. This survey resulted in the discovery of 80 nonprocessed human pseudogenes that have emerged since human and chimpanzees last shared a common ancestor. These genes formally encoded proteins with a number of different functions, but genes involved in chemoreception and immune response are significantly overrepresented in this set. To further explore how gene loss can lead to human-specific adaptations, the authors focus in on one specific example: *CASPASE12*. Based on a population genetic analysis, the authors present evidence that pseudogenization of this locus was driven by positive selection related to protection from sepsis. They were also able to estimate that the process of *CASPASE12* pseudogene fixation started just prior to human migration out of Africa. Consistent with this proposed scenario, two other genes related to sepsis were rendered non-functional along the human lineage. This study underscores the potential importance of gene loss in the evolution of species-specific adaptations.