Genome Analysis

Evolutionary footprints of nucleosome positions in yeast

Stefan Washietl[1,2], Rainer Machné[2] and Nick Goldman[1]

[1] European Molecular Biology Laboratory – European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK
[2] Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria

Using genome-wide maps of nucleosome positions in yeast, we have analyzed the influence of chromatin structure on the molecular evolution of genomic DNA. We have observed, on average, 10–15% lower substitution rates in linker regions than in nucleosomal DNA. This widespread local rate heterogeneity represents an evolutionary footprint of nucleosome positions and reveals that nucleosome organization is a genomic feature conserved over evolutionary timescales.

Background

Eukaryotic genomes are packed into nucleosomes, repeating units of ~147 base pairs of DNA wrapped around a histone protein complex and connected by free linker DNA [1]. Nucleosome organization determines DNA accessibility with important consequences for genome function [2]. Several recent studies have used high-throughput experimental techniques to generate genome-scale maps of nucleosome positions in the yeast *Saccharomyces cerevisiae* [3–7]. A surprisingly large fraction, 70–80%, of the genome was found to be occupied by well-positioned nucleosomes. These data not only provide important new insights into nucleosome organization, but also permit us to address interesting new questions for the first time. Here, we address the question of whether there is a connection between the molecular evolution of genomic DNA and its packaging in the nucleus.

Substitution rates are strongly correlated with nucleosome positions

We analyzed substitution rates in the yeast genome and their dependence on the position within a map of experimentally determined nucleosome locations [5]. From genome-wide alignments [8] of five closely related *Saccharomyces* species of the *sensu stricto* group (*S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*), we extracted columns corresponding to the 147 positions within the nucleosome and ten linker positions before and after. We obtained 167 (10+147+10) sub-alignments of ~59 000 columns from all genomic regions (genic and intergenic). From these alignments, we estimated the branch lengths for the five-species phylogenetic tree [9] using maximum likelihood (Supplementary Methods and Figure S1 in the supplementary material online). We observed a strong correlation between the tree length (used as a measure of substitution rate) and the position within the nucleosome (Figure 1a). This positional dependency is highly significant: the null hypothesis that the observed pattern is produced by chance is rejected with $P{\sim}10^{-25}$ (two-sided Wald–Wolfowitz runs test). Substitution rates are ~10% lower in the linker regions than in the region around the dyad (the equidistant centre point of nucleosomal DNA).

Nucleosome-related rate heterogeneity is found in intergenic and coding regions

The observed differences in substitution rates could be the result of a difference in mutation rate, a difference in selection pressure or a combination of both. Higher levels of selective constraints in linker DNA might be a consequence of regulatory sites being enriched in linker regions that are more accessible for binding factors. A correlation between transcription-factor binding and nucleosome occupancy is well-established and was confirmed by genome-scale nucleosome maps [3,6]. If higher levels of conservation of regulatory binding sites are responsible for the observed nucleosome-related substitution rate heterogeneity, we would expect this effect to be only present in intergenic regions. Therefore, we recalculated rates separately for intergenic regions and for coding regions. For the latter, we also analyzed the three codon positions separately to see how functional sites with different levels of selective constraints are affected. Moreover, to exclude any potentially unforeseen experimental bias in the nucleosome data from Whitehouse *et al.* [5], we added an additional two datasets of nucleosome positions from Lee *et al.* [6] and Shivaswamy *et al.* [7].

A significant positional dependency of the substitution rates in all tested subsets (*P*-values between 10^{-4} and 10^{-20}) was observed. The relative substitution rate difference between linker and dyad is, approximately, equally strong (~10%) in both intergenic and genic regions (Figure 1b). Within the coding regions, the highly constrained first and second positions are as similarly affected as the more variable third position. The results are also consistent between the different experimental datasets. However, in the Shivaswamy *et al.* [7] set, the signal in the coding regions is slightly below the other two sets, possibly indicating a lower accuracy of nucleosome positions. Owing to biological and experimental noise, the discrete nucleosome positions used in our analysis naturally vary in their degree of confidence. A consensus set from all three experiments comprising ~15 000 nucleosome positions

Corresponding author: Washietl, S. (washietl@ebi.ac.uk).

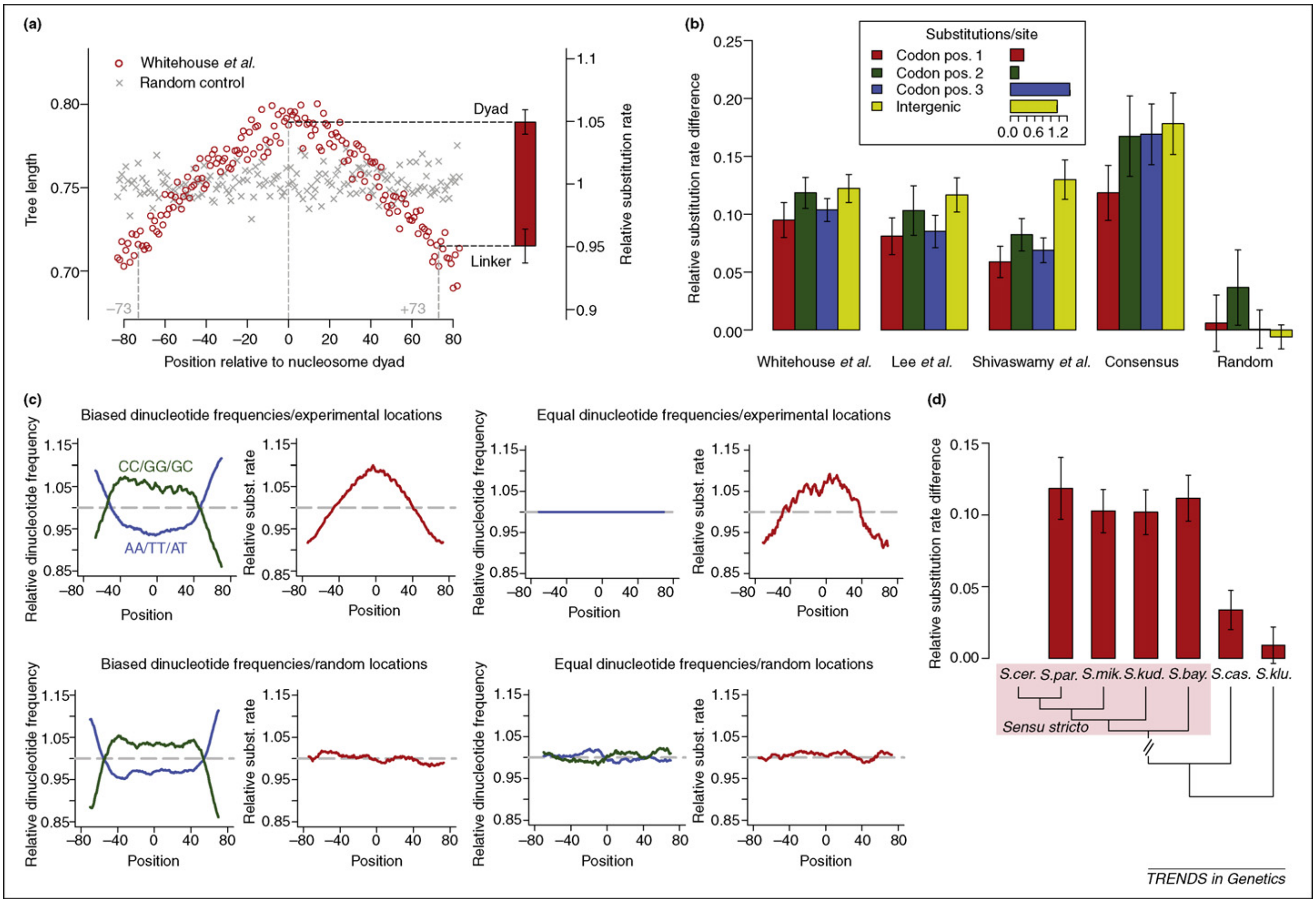


Figure 1. **(a)** Substitution rates of genomic DNA are strongly correlated with their relative position in a map of experimentally determined nucleosome positions. Rate estimates (red circles) are highest around the nucleosome dyad and lowest in the linker, showing a relative difference of ~10% (red bar indicates the difference between the average of 20 points centred around position 0 and 20 points around ± 73). A control, for which position assignments have been randomized, is shown in gray. The vertical axis shows the sum of branch lengths for trees of five yeast species estimated by maximum likelihood. On the right, a normalized (relative) substitution rate is shown by setting the average of the random control to 1. **(b)** Relative substitution rate difference between nucleosome dyad and linker for different experimental datasets and genomic annotations, with ~95% confidence levels for these differences. The inset shows the average (absolute) branch length for the different annotations. 'Random' refers to a control set of randomly selected genomic locations. 'Consensus' is a set consisting of overlapping (and averaged) nucleosome positions from the three different experimental datasets (Supplementary Methods). **(c)** Effect of dinucleotide composition. Nucleosome data show strongly biased dinucleotide frequencies, comparable in strength to the signal in substitution rates (upper left). In a subset of sites with equal dinucleotide content for each position, the substitution rates still show the same positional dependency (upper right). Random locations with biased dinucleotide content (lower left) and completely random locations (lower right) do not show differences in substitution rate. Data are shown for the consensus set (codon position 2) from (a). All curves are a running average over 20 positions. **(d)** Relative substitution rate differences calculated in pairwise comparisons of *S. cerevisiae* to *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castelli* and *S. kluyveri*.

(Supplementary Methods) showed a higher substitution rate difference than any of the single experiments (up to ~15%) (Figure 1b). In addition, we found that the phylogenetic signal directly correlates with the significance scores of inferred nucleosome positions in all three datasets (Figure S2).

Rate heterogeneity is independent of dinucleotide composition

Given the strong signal in coding regions, it is unlikely that regulatory DNA-binding sites are the reason for higher conservation levels in linkers. As an alternative explanation of our observations, we have to consider nucleosome positioning signals in the DNA. Typical ~10 base pair dinucleotide periodicities favouring DNA bendability have been described to be characteristic of nucleosomal DNA [10] and can explain some *in vivo* nucleosome positions in yeast [11,12]. In the much larger genome-wide datasets used here, we could not identify similarly strong periodicities and, also, did not find any related periodicity in the substitution rates. There are also sequence characteristics that inhibit nucleosome formation [6,13–15] and negative selection acting on such 'exclusion signals' might explain lower substitution rates in the linker. Consistent with the study by Peckham *et al.* [14], we observe differences in overall dinucleotide composition between linker and nucleosome. The non-random distribution of the most affected dinucleotides ($P{\sim}10^{-14}$, two-sided runs test) is shown in Figure 1c. This dinucleotide bias might be caused by evolved nucleosome positioning signals in the DNA (Note S1). To see if this potential 'nucleosome code' is related to the observed patterns of substitution rates, we calculated the rates only for a subset of sites that were selected to have exactly the same dinucleotide composition for each position (Figure 1c). We still observe similar differences in substitution rates, indicating that this effect is not caused by a selection constraint acting on specific dinucleotides. This result also rules out the possibility that the substitution rate heterogeneity is only an indirect effect of the dinucleotide content, which causes differences in the substitution rate by itself. As an additional control, we randomly placed nucleosomes in the genome with a bias favouring location that matched the dinucleotide profile of the real data (Supplementary Methods). This resulted in a set of similarly biased average dinucleotide content but showed no difference in the substitution rate (Figure 1c).

Substitution rate footprints indicate evolutionary conservation of nucleosome positions

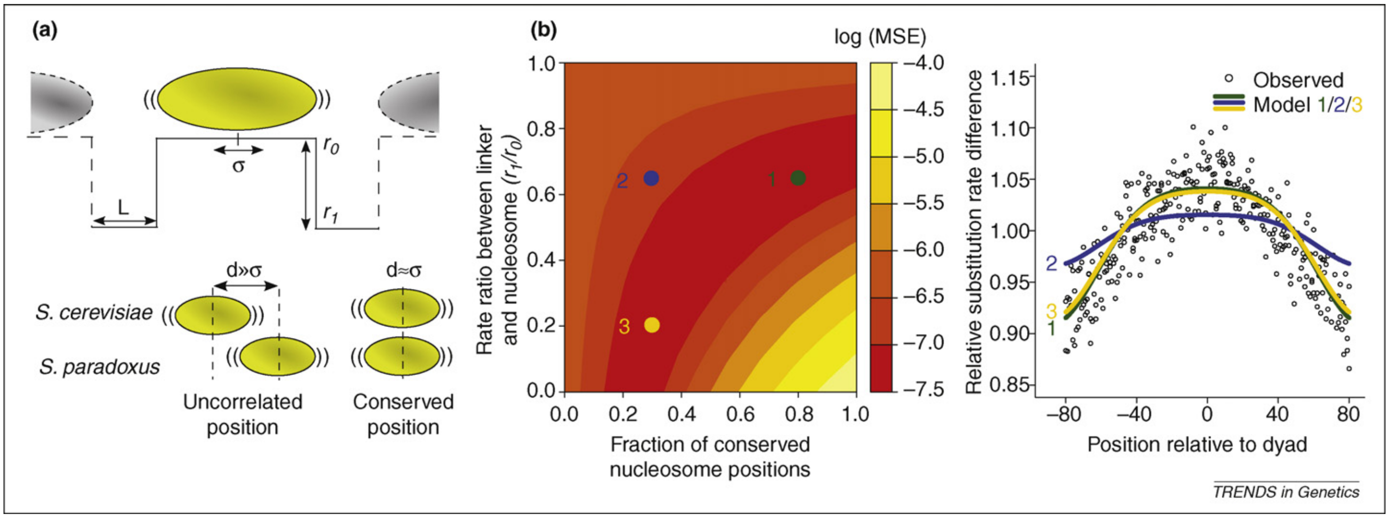
The fact that the nucleosome-related rate heterogeneity affects all genomic sites independently of genomic annotation, selection constraints or base composition implies that this effect is the result of a general difference in mutation rate. Initially, it seems counter-intuitive that 'naked' linker DNA is more conserved than 'protected' nucleosome DNA. Interestingly, a multitude of experimental studies show that nucleosomes form a strong barrier for DNA repair proteins, often resulting in higher repair efficiencies in linker regions [16,17]. This is particularly well-established for repair mechanisms of UV lesions by the photoreactivation pathway [18] and the nucleotide-excision pathway [19]. For the base excision repair pathway, which targets damage from oxidation or alkylation, it has also been shown that naked DNA is more efficiently repaired than nucleosomal DNA [20]. For the mismatch repair pathway, which corrects replication errors, less is known in this context, but a connection between its efficiency and chromatin structure has been suggested [21].

Although efficient DNA-damage repair in linkers would be a very intuitive explanation for our observations, many other cellular and molecular factors could be involved. A phylogenetic study cannot address the mechanistic reasons for the observed differences, but the signal *per se* has some interesting implications.

The fact that we can observe a substitution rate difference when comparing sequences of contemporary species that diverged ~20 million years ago necessarily implies that nucleosome positions are evolutionarily conserved. If nucleosome positions changed randomly over time, the observed substitution rate would average out at a mean level and no 'footprints' would be visible. However, it cannot be assumed that all positions are perfectly conserved, even in closely related species. In principle, it is possible that only a few well-conserved positions cause the phylogenetic signal.

We used a theoretical model to study how nucleosome positioning in two species can give rise to the observed substitution rate patterns (Figure 2a and Supplementary Methods). The model assumes two rates for linker and nucleosomal DNA and takes into account the inherent uncertainty of the experimental positions (Figure 2a). It further considers conserved nucleosomes that have fixed locations and nucleosomes with uncorrelated positions (Figure 2a). We calculated the quality of the fit of the theoretical model to the observed data depending on two factors: (i) the fraction of conserved nucleosome positions; and (ii) the actual substitution rate difference between nucleosomal and linker DNA. Assuming, for example, a rate ratio between nucleosome and linker of 0.65, approximately 80% of the nucleosomes need to be conserved to get an optimal fit (Figure 2b). If only 30% were conserved at this rate ratio, there is no way to reach the observed rate difference and the fit is very poor (Figure 2b). In principle, our model could explain the data with only 30% of conserved nucleosomes (Figure 2b). However, this would require rate ratios of approximately 0.2 (i.e. differences as high as fivefold, which is even higher than the typical substitution rate differences between coding and noncoding regions). We cannot rule out this scenario, but it seems unrealistic because it would indicate that the effect of chromatin structure on the substitution rate is stronger than one of the currently strongest selective forces known in the yeast genome [22].

Although this model is based on some simplified assumptions, it captures three important aspects of the



TRENDS in Genetics

Figure 2. **(a)** A simple model can explain the observed patterns of rate heterogeneity. We assume two different rates for nucleosomal DNA (r_0) and linker DNA (r_1), a level of uncertainty σ for the nucleosome positions (being a result of experimental and biological noise) and an average linker length *L*. When comparing positions in two species, we assume two extreme scenarios: (i) either two nucleosomes have uncorrelated positions (the distance variation *d* between the species is much higher than σ); or (ii) they have conserved positions (the difference between species is approximately the same as σ). **(b)** Fit of the theoretical model to the observed rates between *S. cerevisiae* and *S. paradoxus*. The quality of the fit depends on the rate ratio and the fraction of conserved nucleosome positions (heatmap; shown on the left). On the right, model predictions are shown for three parameter pairs together with the real data. Abbreviations: MSE, mean squared error; note the logarithmic scale.

problem. It shows that: (i) the pattern of the positional dependency can be approximated by assuming only two distinct rates; (ii) the actual rate difference is likely to be higher than the observed difference; and (iii) the observed rate difference cannot be explained without a substantial fraction of conserved nucleosomes positions.

To test how deep in the phylogeny the mutation rate footprints are present, we calculated substitution rates in pairwise comparisons of *S. cerevisiae* to the other four species. We also included the two more distantly related species, *Saccharomyces castelli* and *S. kluyveri*, from the *sensu lato* and petite negative groups, respectively (Figure 1d). The pairwise rate difference with the four species from the *sensu stricto* group is approximately the same as in the five-way comparison. However, no significant signal is seen in the comparison to the two more distantly related species ($P{\sim}0.07$ and $P{\sim}0.30$, for *S. castelli* and *S. kluyveri*, respectively; two-sided runs test), indicating that at this greater evolutionary divergence (presumably >100 million years) the majority of nucleosome positions have changed relative to the underlying DNA sequence. This result is consistent with the view that nucleosome organization is tightly linked to gene organization, which is very similar in the *Saccharomyces* species of the *sensu stricto* group but has undergone major changes in the more distant species [23].

To our knowledge, nucleosome-related substitution rate heterogeneity has not been studied before. It has, however, been indirectly indicated in a previous study reporting periodicity of single nucleotide polymorphisms around transcription start sites in human transcripts [24]. Using nucleosome position data from different human cell-lines [25,26] and alignments of human, chimp and macaque genomes, we investigated whether there is a correlation between nucleosome positions and substitution rates in primates. We could not identify rate differences comparable to the signal in yeast (Note S2, Figure S4). Unlike yeast, only a rather small fraction of the large human genome seems to be covered by well-positioned nucleosomes. Given that nucleosome positioning is clearly linked to gene expression [25], variations in different cell lines and developmental stages must be considered. A mutation rate 'footprint' would not reflect a global average nucleosomal state of the genome but only a specific situation in the germ-line. In this light, yeast, a unicellular eukaryotic organism with a compact genome, seems to be an ideal model organism to study this effect.

Concluding remarks

We have described a direct connection between nucleosome positions and substitution rates of genomic DNA in yeast. Using the available data, we can estimate a lower bound of 10–15% difference of substitution rates between nucleosomal and linker DNA. This impact of chromatin structure on the evolutionary processes of genomic DNA is not only statistically highly significant, it is also of such a magnitude that it needs to be considered as an important factor shaping the genomic landscape. Interpreted as a 'footprint' of nucleosomes, this rate heterogeneity enables an evolutionary analysis of nucleosome positions using techniques from molecular phylogenetics. These footprints seem to be independent of previously described positioning signals and clearly show that nucleosome organization is a feature of the yeast genome that is conserved over evolutionary timescales.

However, it is necessary to investigate in more detail the causes of the nucleosome-related rate heterogeneity, especially to find out the degree to which selective constraints and mutation rate differences are involved. If the latter can be confirmed as one of the main forces responsible for the observed footprints, it will be of particular interest to identify the corresponding molecular or cellular mechanisms.

Update

After this article was accepted for publication, we learned that Warnecke *et al*. [27] have reported similar findings to ours. The authors of this study favour selective constraints from nucleosome positioning signals as an explanation for low substitution rates in linker regions.

Acknowledgements

S.W. is supported by a fellowship of the Austrian genome research program GEN-AU. R.M. acknowledges financial support by WWTF project number MA07–030. We thank Ari Löytynoja and Martin Taylor for comments on the manuscript and Dustin Schones for providing assistance and software to analyze the human dataset.

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2008.09.003.

References

1 Richmond, T.J. and Davey, C.A. (2003) The structure of DNA in the nucleosome core. *Nature* 423, 145–150
2 Rando, O.J. and Ahmad, K. (2007) Rules and regulation in the primary structure of chromatin. *Curr. Opin. Cell Biol.* 19, 250–256
3 Yuan, G.C. *et al.* (2005) Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* 309, 626–630
4 Albert, I. *et al.* (2007) Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* 446, 572–576
5 Whitehouse, I. *et al.* (2007) Chromatin remodelling at promoters suppresses antisense transcription. *Nature* 450, 1031–1035
6 Lee, W. *et al.* (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.* 39, 1235–1244
7 Shivaswamy, S. *et al.* (2008) Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.* 6, e65
8 Kuhn, R.M. *et al.* (2007) The UCSC genome browser database: update 2007. *Nucleic Acids Res.* 35, D668–D673
9 Rokas, A. *et al.* (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, 798–804
10 Satchwell, S.C. *et al.* (1986) Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* 191, 659–675
11 Segal, E. *et al.* (2006) A genomic code for nucleosome positioning. *Nature* 442, 772–778
12 Ioshikhes, I.P. *et al.* (2006) Nucleosome positions predicted through comparative genomics. *Nat. Genet.* 38, 1210–1215
13 Iyer, V. and Struhl, K. (1995) Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.* 14, 2570–2579
14 Peckham, H.E. *et al.* (2007) Nucleosome positioning signals in genomic DNA. *Genome Res.* 17, 1170–1177
15 Yuan, G.C. and Liu, J.S. (2008) Genomic sequence is highly predictive of local nucleosome depletion. *PLOS Comput. Biol.* 4, e13

16 Thoma, F. (2005) Repair of UV lesions in nucleosomes–intrinsic properties and remodeling. *DNA Repair (Amst.)* 4, 855–869

17 Ataian, Y. and Krebs, J.E. (2006) Five repair pathways in one context: chromatin modification during DNA repair. *Biochem. Cell Biol.* 84, 490–504

18 Suter, B. and Thoma, F. (2002) DNA-repair by photolyase reveals dynamic properties of nucleosome positioning *in vivo*. *J. Mol. Biol.* 319, 395–406

19 Ura, K. *et al.* (2001) ATP-dependent chromatin remodeling facilitates nucleotide excision repair of UV-induced DNA lesions in synthetic dinucleosomes. *EMBO J.* 20, 2004–2014

20 Beard, B.C. *et al.* (2003) Suppressed catalytic activity of base excision repair enzymes on rotationally positioned uracil in nucleosomes. *Proc. Natl. Acad. Sci. U. S. A.* 100, 7465–7470

21 Hawk, J.D. *et al.* (2005) Variation in efficiency of DNA mismatch repair at different sites in the yeast genome. *Proc. Natl. Acad. Sci. U. S. A.* 102, 8639–8643

22 Siepel, A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050

23 Cliften, P. *et al.* (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* 301, 71–76

24 Higasa, K. and Hayashi, K. (2006) Periodicity of SNP distribution around transcription start sites. *BMC Genomics* 7, 66

25 Schones, D.E. *et al.* (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell* 132, 887–898

26 Ozsolak, F. *et al.* (2007) High-throughput mapping of the chromatin structure of human promoters. *Nat. Biotechnol.* 25, 244–248

27 Warnecke, T. *et al.* (2008) The impact of the nucleosome code on protein-coding sequence evolution in yeast. *PLoS Genet.* (in press)

0168-9525/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2008.09.003 Available online 24 October 2008

Genome Analysis

Inferring molecular function: contributions from functional linkages

Arturo Medrano-Soto[1,2], Debnath Pal[3] and David Eisenberg[1,2]

[1] Howard Hughes Medical Institute (HHMI), 675C. E. Young Drive South, Los Angeles, CA 90095, USA
[2] UCLA-DOE Institute for Genomics and Proteomics, 611C. E. Young Drive East, 201 Boyer Hall, Los Angeles, CA 90095, USA
[3] Bioinformatics Center and Supercomputer Education Research Center, C.V. Raman Circle, Indian Institute of Science, Bangalore 560012, Karnataka, India

In the current era of high-throughput sequencing and structure determination, functional annotation has become a bottleneck in biomedical science. Here, we show that automated inference of molecular function using functional linkages among genes increases the accuracy of functional assignments by $\geq$8% and enriches functional descriptions in $\geq$34% of top assignments. Furthermore, biochemical literature supports >80% of automated inferences for previously unannotated proteins. These results emphasize the benefit of incorporating functional linkages in protein annotation.

Functional linkages and annotation of protein function

The current flood of complete genome sequences, coupled with the substantial progress of structural genomics, has deluged scientists with myriad protein sequences and structures for which there is often little or no functional information. This flood of data has stimulated the development of a body of computational methods to reveal the likely biological roles of unannotated proteins (for recent reviews, see Refs [1–4]). Functional linkages – genes identified as functionally related by bioinformatic approaches based on genomic context – have mainly been used to gain insights into the cellular processes in which genes participate [5,6]; for instance, in model organisms *Escherichia coli* K12 and *Bacillus subtilis*, ~70% of all pairs of genes within operons share similar biological processes (see Supplementary Material online). However, little attention has been devoted to learning how these relationships might contribute to the specific task of inferring molecular function. A preliminary estimate of the utility of functional linkages is available from the observation that, in *E. coli* K12 and *B. subtilis*, >40% of gene pairs within operons share very similar molecular functions (see Supplementary Material online). This indicates that computational methods aiming to infer or assign a molecular function to a protein can benefit from a better understanding of functional linkages. Here, we use the ProKnow metaserver (http://proknow.mbi.ucla.edu) [7] (Box 1) as a tool to assess the extent to which the quality of assignment of molecular function can be improved by incorporating annotations collected from proteins functionally linked to the query protein. We believe this work is the first attempt to quantify the contribution of information on functional linkages to the inference of molecular function.

Assessing the contribution of functional linkages to inference of molecular function

We have added a new feature extractor to the ProKnow metaserver, whereby the Gene Ontology (GO; http://www.geneontology.org/) annotations of proteins that are inferred to be functionally linked to the query by methods based on genomic context available in the ProLinks database (http://prolinks.mbi.ucla.edu) [8] are taken into account in the inference process (Box 1). Hereafter, we refer to this feature extractor as the ProLinks module. We evaluated functional assignments using two test sets of proteins extracted from the Protein Data Bank (PDB; http://www.rcsb.org/pdb/) [9] based on protein sequence identity and type of fold. The first set consisted of 599 representative PDB proteins showing <50% sequence

Corresponding author: Eisenberg, D. (david@mbi.ucla.edu).