6.047 / 6.878 Computational Biology: Genomes, Networks, Evolution
Fall 2008

Genome wide study of gene regulation

6.047/6.878 Lecture 20

Lectured by P. Kheradpour, 11/13/08

Table of contents

Gene expression patterns (motivation)

Given the complex spatial and temporal regulation of gene expression [7], exemplified by recent studies on the mouse [21], and on Drosophila heart development [30], there is great impetus to identify the targets of regulatory factors in order to understand how such regulation is achieved. Many connections in these regulatory networks are conserved all the way up to mammmals, and so we can conceivably use evolutionary information in order to probe the regulatory networks of humans.

We will focus here on regulation of transcription – as achieved by *transcription factors* (TFs) – and also regulation of translation by microRNAs (miRNAs).

Transcription factors

TFs regulate the expression of target genes by binding to DNA, in regions that may or may not be proximal to the gene that is being regulated. Binding is specific for a particular target sequence, or *motif*, although there is most probably also non-specific binding that enables the TFs to bind to other regions of DNA in order to rapidly locate their target sequences [10]. Each motif can be viewed as a position weight matrix (PWM), which shows the relative specificity for each base at every position in the motif. The effect of TF binding may be to activate or to repress gene expression.

We will now examine some of the approaches that have been used to discover motifs.

Experimental discovery

Several revolutionary experimental methods have enabled motif discovery to be carried out in a brute force, *de novo* fashion, without need for knowledge of promoter regions.

The **S**ystematic **E**volution of **L**igands by **Ex**ponential Enrichment (SELEX) protocol [32] involves taking a pool of RNA or DNA fragments (the *library*) and adding the protein of interest. The members of the library that do not bind to the protein are then removed, and another batch of fragments added. This results in enrichment of the sequences that bind favourably to the protein, although it does not allow particularly for the binding strength to be quantified.

The DNA immunoprecipitation with microarray detection (DIP-Chip) approach [16] involves adding naked genomic DNA fragments of around 600bp to a sample of purified protein. The protein-bound fragments are separated and labelled, such that they can be easily identified by binding to a whole-genome microarray.

Protein binding microarrays (PBMs) [20] contain double stranded DNA, meaning that the protein of interest can bind directly to the microarray. By tagging the protein with a dye, or antibody recognition sequence, the PBM can be used to visualize which sequences bind the protein. By spotting all possible k-mers on a given chip, a relative intensity can be calculated for each sequence, allowing the binding strength to be quantified.

Computational discovery

The first approach is to take coregulated genes and perform some kind of enrichment process. This enrichment may be achieved by determining which motifs occur in at least n sequences [23], or through approaches such as EM or Gibbs sampling (*see Lecture 9*) [17, 31].

Another approach to motif discovery is to look for conservation in multiple genomes. Given the conservation rate for random motifs, it is possible to assign a p-value to such conserved regions [33].

This process was pipelined by Kellis *et al.* [11] in the following fashion:

i) create motif seeds of the form `XXX-(--)`$_n$`-XXX` using six non-degenerate characters (`ACGT`)

– note: motifs often have a gap in the middle (around 0-10bp), hence the inclusion of a gap in the seeds. A length-10 gap may be biologically justified since this is one turn of the helix, and so this motif will correspond to two distinct binding sites on the same side of DNA

ii) score conservation

iii) expand seeds using degenerate characters

iv) cluster motifs using sequence similarity

Discovered motifs have functional enrichments

Using a similar procedure to that detailed above, an analysis of 12 Drosophila genomes [29] showed that certain motifs were enriched in particular tissues (*red dots in figure on slide 14*), and that functional clusters of similar motif expression patterns emerged in related tissues. Conversely, ubiquitously expressed genes showed a particularly low occurrence of most motifs, indicating that these genes are not regulated by TFs.

The earlier studies on mammals [33] showed positional bias with respect to the transcription start site (TSS), and tissue-specific expression for many of the computationally discovered motifs, all of which suggests that the computational procedure is detecting motifs that are actually involved in regulation.

Summary

Despite numerous successful applications, experimental approaches are costly and slow, and generally require purified TF. It may also be necessary to apply computational procedures to process the microarray data obtained, so one cannot regard them as purely experimental techniques.

However, computational approaches also require sets of coregulated genes or multiple genomes in order to carry out the procedures outlined above. In addition, whilst it is not necessary to possess *a priori* knowledge of the TFs before carrying out these analyses, this can mean that it is difficult to subsequently match the discovered motifs to specific TFs.

microRNAs

The biology

microRNAs are short pieces of DNA around 20 nucleotides in length, which are transcribed from the genome by Pol II. After several processing steps, including by the nuclease Drosha, the miRNA is exported from the nucleus. It then has the loop removed by an enzyme known as Dicer, leaving an miRNA:miRNA duplex. This is eventually cleaved by a helicase, and one section of single-stranded RNA then joins with the RNA-induced silencing complex (RISC), forming the miRISC [3]. This complex binds to a target mRNA sequence, and disrupts translation by interrupting the progress of the ribosomes. The miRISC also marks the mRNA for degradation.

Whilst only discovered fifteen years ago [13], miRNAs are now known to be involved in almost every level of regulation, and most genes are now thought to be targetted by miRNAs [9]. In addition to the mechanisms of translation regulation already mentioned, it has been suggested that the miRNAs may help to filter out noise in expression levels [28], as well as cleaning up residual mRNAs.

The structure of the miRNA tends to be a stem-loop, with two complementary arms linked by a loop region [3]. One of these arms contains a binding sequence that is complementary to the target mRNA, and this arm will become part of the miRISC. Occasionally the other arm may also bind a different mRNA sequence, in which case this second arm is termed the *star* arm. Other structural features may include small loops along the arms where there is mismatch in complementarity. Interestingly, despite the length of the miRNA sequence, the specificity of the miRNA for its target mRNA sequence is conferred mainly by a 7-8 nucleotide stretch at the start of the region of the binding arm of the miRNA that goes on to form the miRISC.

Computational miRNA discovery

Given the distinctive structural features of miRNAs, one computational strategy for detecting them is to search the genome for likely hairpins, featuring two complementary regions with a short loop in between. However, many such hairpins exist, and very few true miRNAs are among these. As such, very high specificity is needed for reliable prediction.

It has been observed that structural features alone are insufficient to reliably predict miRNAs, even when folding energy and other features are all incorporated into a machine learning approach. Screening for hairpin energy does result in around a 40-fold enrichment for true miRNAs, but this is not sufficient given the huge number of putative hairpins.

12 Drosophila genomes

Using the alignment of the genomes of twelve species of Drosophila from Clark *et al.* [1], it was observed that the miRNA arms are highly conserved, whilst the loop tends to be less conserved. Given that one of the arms actually binds to the mRNA, whereas the loop is cleaved and has no specific function (in these examples at least), one might expect this pattern of conservation. It was also noted that the regions flanking the loop are not well conserved, and these different patterns of conservation give rise to a characteristic conservation profile for miRNAs.

Using a machine learning approach on this conservation profile, Ruby *et al.* [24] were able to achieve much higher enrichment for true miRNAs. Indeed, combining this with all the other features enabled a 4551-fold enrichment, which could be deemed sufficient for predictive purposes.

Validation of the predicted miRNAs

Experimental approaches have been used to sequence miRNAs (usually just the arms), and a the predicted miRNAs were observed to show a good match with these experimental reads [24]. Of the 101 hairpins obtained through the computational procedure, good evidence was found to suggest that most of these were in fact true miRNAs. Of those that did not match any existing evidence, many were found within introns and intergenic regions, which indicates that the predictions are in fact reasonable. In addition, many predicted miRNAs were seen to cluster with other known miRNAs.

In theory, both sense and anti-sense strands of the miRNA gene could fold up to form the same hairpin, with the star arm of the sense hairpin equivalent to the non-star arm of the antisense hairpin [24]. Nevertheless, there was an observed preference for the sense strand in many cases [29], although this could provide an additional mechanism for variability.

Two predicted miRNAs were observed to overlap with protein coding genes [15], and as such these were initially discarded. However, as it turned out, these regions of the genome had in fact been erroneously annotated as protein coding regions due to the transcription of the miRNA genes.

Additional signatures of mature miRNAs

As already mentioned, the 7mer at the 5' of the mature miRNA (or miRNA*) sequence is chiefly responsible for the specificity of binding. This short region was observed to be highly conserved, particularly in the non-star sequence, and as such, the high conservation of this region can be used to identify the adjacent cleavage site, and to predict the 5' end of the miRNA. It was also observed that this 7mer tends to be conserved in the 3' UTRs of target sequences, but avoided in those of anti-targets, demonstrating both positive and negative selection.

In the cases where it was not possible to predict the 5' end accurately, imprecise processing was observed to be associated with these miRNAs [29], providing another example of the link between the computational and biological features of miRNA regulation.

Question: how much of miRNA specificity comes from the RISC complex?

Answer: It is mostly bases 2-7 of the miRNA that determine specifity, with the RISC acting as more of a scaffold. The first base of the 7mer is less important, and tends to be a T or a U. There is in fact a crystal structure of an miRNA:mRNA complex from *C.elegans* that was published earlier this year [6].

Prediction of binding sites

With motif discovery one can look for statistical enrichment over the whole genome in order to establish the consensus for the motif. However, in order to find a single instance of a binding site (whether for a TF binding site, or an miRNA target), different procedures must be adopted [12].

Experimental approaches

Several chromatin immunoprecipitation (ChIP) techniques have been developed in order to detect protein-DNA binding. The general approach consists of using antibodies for a particular protein to immunoprecipitate sequences bound to the protein. This can be followed by either microarray detection (ChIP-Chip), or sequencing (ChIP-Seq) [18] of the bound sequences.

However, both of these approaches rely on antibody availability, and are restricted to specific tissues. Although this specificity can also be useful, the experimental approaches are plagued by high false positive and false negative rates. In principle these approaches find all binding sites, but if there is lots of TF then low-affiinity sites may become bound, and so the definition of a binding site will depend on the specific conditions used.

Computational approaches

The main role for computational approaches is to increase the specificity of matching. Single-genome approaches may look for clustering of binding sites for TFs that are thought to act together [22, 4, 27], but these techniques may miss instances where TFs act alone. Multi-genome approaches, such as phylogenetic footprinting [19, 5, 8, 14] are therefore of interest.

Phylogenetic footprinting

Within a given multiple sequence alignment (MSA), there may be regions that are not present in all the aligned sequences. In particular, some binding sites for a specific motif may only be conserved in a subset of the species, and may have moved, or mutated in other species. Traditional conservation-based approaches to detecting binding sites might therefore miss these cases.

One possible way of tackling this problem is to use the notion of the branch length score (BLS) to assign a score to particular instances where a binding site is 'missing' in certain sequences. The BLS is computed by taking the total length of all the branches in the phylogenetic tree connecting the sequences that possess the binding sequence, and dividing this total by the sum of the branch lengths for the tree connecting all sequences [12]. The result of this is that the absence of the binding sequence from a few short branches of the tree does not have a large impact on the resulting score.

The BLS can be converted into a confidence score by using random motifs to work out the average background noise, and then expressing the confidence score as $C = \frac{\text{signal}}{\text{signal} + \text{noise}}$.

This confidence score has been shown to select for TFs that occur within promoters, and miRNA motifs in the 3' UTRs of their target mRNAs

Confidence score must be saying something, since high confidence score enriches for strand bias in miRNAs, as well as recapitulating the strand bias observed in miRNA motifs.

Comparison with experiment

When the high-confidence binding sites are compared to sequences derived from ChIP data, a high enrichment for ChIP sequences is observed in both mammals and flies [2]. If the binding sequences are highly conserved as well as posessing a high confidence score, an even higher enrichment is observed. When the promoters of fly muscle genes were examined, there is a clear enrichment for activating motifs in genes that are highly expressed in muscle tissue, and a corresponding enrichment for repressor motifs in underexpressed genes. The computational procedure was able to detect this enrichment better than the ChIP data [34, 26, 25] in almost all cases.

It is worth noting, however, that an advantage of the ChIP experiments over computational procedures that employ evolutionary information is that the former allow detection of species-specific binding, whereas the latter rely on a signal across species.

Network level examination

The regulatory network for the fly was reconstructed using the computational predictions, and was seen to feature several links that are confirmed by the literature. It was also observed that many genes are highly regulated by both TFs and miRNAs, as shown by the correlated in-degrees in the network. All of this could be taken to support the validity of the computational procedures that have been described.

Bibliography

[1] Evolution of genes and genomes on the drosophila phylogeny. *Nature*, 450(7167):203–218, November 2007.

[2] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823 – 837, 2007.

[3] David P. Bartel. Micrornas: Genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281 – 297, 2004.

[4] Benjamin P. Berman, Yutaka Nibu, Barret D. Pfeiffer, Pavel Tomancak, Susan E. Celniker, Michael Levine, Gerald M. Rubin, and Michael B. Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):757–762, 2002.

[5] Mathieu Blanchette and Martin Tompa. Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting. *Genome Research*, 12(5):739–748, 2002.

[6] M. Cevec, C. Thibaudeau, and J. Plavec. Solution structure of a let-7 mirna:lin-41 mrna complex from c. elegans. *Nucleic Acids Research*, 36(7):2330–2337, February 2008.

[7] Eric H. Davidson and Douglas H. Erwin. Gene Regulatory Networks and the Evolution of Animal Body Plans. *Science*, 311(5762):796–800, 2006.

[8] Laurence Ettwiller, Benedict Paten, Marcel Souren, Felix Loosli, Jochen Wittbrodt, and Ewan Birney. The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. *Genome Biology*, 6(12):R104, 2005.

[9] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, In press, 2008.

[10] Stephen E. Halford and John F. Marko. How do site-specific DNA-binding proteins find their targets?⟨⟩ *Nucl. Acids Res.*, 32(10):3040–3052, 2004.

[11] Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, May 2003.

[12] Pouya Kheradpour, Alexander Stark, Sushmita Roy, and Manolis Kellis. Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Research*, 17(12):1919–1931, 2007.

[13] R. C. Lee, R. L. Feinbaum, and V. Ambros. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *Cell*, 75(5):843–854, December 1993.

[14] B. P. Lewis, I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge. Prediction of mammalian microrna targets. *Cell*, 115(7):787–798, December 2003.

[15] Michael F. Lin, Joseph W. Carlson, Madeline A. Crosby, Beverley B. Matthews, Charles Yu, Soo Park, Kenneth H. Wan, Andrew J. Schroeder, L. Sian Gramates, Susan E. St. Pierre, Margaret Roark, Kenneth L. Wiley, Rob J. Kulathinal, Peili Zhang, Kyl V. Myrick, Jerry V. Antone, Susan E. Celniker, William M. Gelbart, and Manolis Kellis. Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. *Genome Research*, 17(12):1823–1836, 2007.

[16] Xiao Liu, David M. Noll, Jason D. Lieb, and Neil D. Clarke. DIP-chip: Rapid and accurate determination of DNA-binding specificity. *Genome Research*, 15(3):421–427, 2005.

[17] Kenzie D MacIsaac and Ernest Fraenkel. Practical strategies for discovering regulatory dna sequence motifs. *PLoS Comput Biol*, 2(4):e36, Apr 2006.

[18] Elaine R. Mardis. Chip-seq: welcome to the new frontier. *Nature Methods*, 4(8):613–614.

[19] A. M. Moses, D. Y. Chiang, D. A. Pollard, V. N. Iyer, and M. B. Eisen. Monkey: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol*, 5(12), 2004.

[20] Sonali Mukherjee, Michael F. Berger, Ghil Jona, Xun S. Wang, Dale Muzzey, Michael Snyder, Richard A. Young, and Martha L. Bulyk. Rapid analysis of the dna-binding specificities of transcription factors with dna microarrays. *Nature Genetics*, 36(12):1331+, November 2004.

[21] L. A. Pennacchio, N. Ahituv, A. M. Moses, S. Prabhakar, M. A. Nobrega, M. Shoukry, S. Minovitsky, I. Dubchak, A. Holt, K. D. Lewis, I. Plajzer-Frick, J. Akiyama, S. De Val, V. Afzal, B. L. Black, O. Couronne, M. B. Eisen, A. Visel, and E. M. Rubin. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502, Nov 23 2006.

[22] Anthony A. Philippakis, Brian W. Busser, Stephen S. Gisselbrecht, Fangxue S. He, Beatriz Estrada, Alan M. Michelson, and Martha L. Bulyk. Expression-guided in silico evaluation of candidate cis regulatory codes for drosophila muscle founder cells. *PLoS Computational Biology*, 2(5):e53+, May 2006.

[23] Isidore Rigoutsos. Combinatorial pattern discovery in biological sequences: the teiresias algorithm. *Bioinformatics*, 14:55–67(13), February 1998.

[24] J. Graham Ruby, Alexander Stark, Wendy K. Johnston, Manolis Kellis, David P. Bartel, and Eric C. Lai. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Research*, 17(12):1850–1864, 2007.

[25] Thomas Sandmann, Charles Girardot, Marc Brehme, Waraporn Tongprasit, Viktor Stolc, and Eileen E.M. Furlong. A core transcriptional network for early mesoderm development in Drosophila melanogaster. *Genes Development* , 21(4):436–449, 2007.

[26] Thomas Sandmann, Lars J. Jensen, Janus S. Jakobsen, Michal M. Karzynski, Michael P. Eichenlaub, Peer Bork, and Eileen E.M. Furlong. A temporal map of transcription factor activity: Mef2 directly regulates target genes at all stages of muscle development. *Developmental Cell*, 10(6):797 – 807, 2006.

[27] Mark D. Schroeder, Michael Pearce, John Fak, Hongqing Fan, Ulrich Unnerstall, Eldon Emberly, Nikolaus Rajewsky, Eric D. Siggia, and Ulrike Gaul. Transcriptional control in the segmentation gene network of drosophila. *PLoS Biology*, 2(9):e271+, September 2004.

[28] Alexander Stark, Julius Brennecke, Natascha Bushati, Robert B B. Russell, and Stephen M M. Cohen. Animal microrna confer robustness to gene expression and have a significant impact on 3'utr evolution. *Cell*, December 2005.

[29] Alexander Stark, Michael F. Lin, Pouya Kheradpour, Jakob S. Pedersen, Leopold Parts, Joseph W. Carlson, Madeline A. Crosby, Matthew D. Rasmussen, Sushmita Roy, Ameya N. Deoras, Graham J. Ruby, Julius Brennecke, Harvard F. Curators, Berkeley, Emily Hodges, Angie S. Hinrichs, Anat Caspi, Benedict Paten, Seung-Won Park, Mira V. Han, Morgan L. Maeder, Benjamin J. Polansky, Bryanne E. Robson, Stein Aerts, Jacques van Helden, Bassem Hassan, Donald G. Gilbert, Deborah A. Eastman, Michael Rice, Michael Weir, Matthew W. Hahn, Yongkyu Park, Colin N. Dewey, Lior Pachter, James W. Kent, David Haussler, Eric C. Lai, David P. Bartel, Gregory J. Hannon, Thomas C. Kaufman, Michael B. Eisen, Andrew G. Clark, Douglas Smith, Susan E. Celniker, William M. Gelbart, and Manolis Kellis. Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature*, 450(7167):219–232.

[30] Pavel Tomancak, Amy Beaton, Richard Weiszmann, Elaine Kwan, ShengQiang Shu, Suzanna Lewis, Stephen Richards, Michael Ashburner, Volker Hartenstein, Susan Celniker, and Gerald Rubin. Systematic determination of patterns of gene expression during drosophila embryogenesis. *Genome Biology*, 3(12):1–14, 2002. This article is part of a series of refereed research articles from Berkeley Drosophila Genome Project, FlyBase and colleagues, describing Release 3 of the Drosophila genome, which are freely available at http://genomebiology.com/drosophila/.

[31] Martin Tompa, Nan Li, Timothy L. Bailey, George M. Church, Bart D. De Moor, Eleazar Eskin, Alexander V. Favorov, Martin C. Frith, Yutao Fu, James J. Kent, Vsevolod J. Makeev, Andrei A. Mironov, William S. Noble, Giulio Pavesi, Graziano Pesole, Mireille Régnier, Nicolas Simonis, Saurabh Sinha, Gert Thijs, Jacques v. van Helden, Mathias Vandenbogaert, Zhiping Weng, Christopher Workman, Chun Ye, and Zhou Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23(1):137–144, January 2005.

[32] C Tuerk and L Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249(4968):505–510, 1990.

[33] Xiaohui Xie, Jun Lu, E. J. Kulbokas, Todd R. Golub, Vamsi Mootha, Kerstin Lindblad-Toh, Eric S. Lander, and Manolis Kellis. Systematic discovery of regulatory motifs in human promoters and 3‘ utrs by comparison of several mammals. *Nature*, aop(434):338–345, March 2005.

[34] J. Zeitlinger, R. P. Zinzen, A. Stark, M. Kellis, H. Zhang, R. A. Young, and M. Levine. Whole-genome chip-chip analysis of dorsal, twist, and snail suggests integration of diverse patterning processes in the drosophila embryo. *Genes Dev*, 21(4):385–390, February 2007.