Journal of Bioinformatics and Computational Biology
Vol. 4, No. 5 (2006) 981–998
© Imperial College Press

COMPARATIVE GENOMICS OF TRANSCRIPTIONAL REGULATION IN YEASTS AND ITS APPLICATION TO IDENTIFICATION OF A CANDIDATE ALPHA-ISOPROPYLMALATE TRANSPORTER

GALINA YU. KOVALEVA

Department of Bioengineering and Bioinformatics
Moscow State University, Moscow, Russia

Institute for Information Transmission Problems, Moscow, Russia

127994, GSP-4, Russia, Moscow, Bol'shoi Karetnyi per., 19
kovaleva@iitp.ru

GEORGII A. BAZYKIN

Department of Ecology and Evolutionary Biology
Princeton University, Princeton, New Jersey, USA
gbazykin@Princeton.EDU

MICHAEL BRUDNO

Department of Computer Science and Banting
& Best Department of Medical Research
University of Toronto, Toronto, Canada
brudno@cs.toronto.edu

MIKHAIL S. GELFAND

Department of Bioengineering and Bioinformatics
Moscow State University, Moscow, Russia

Institute for Information Transmission Problems, Moscow, Russia
gelfand@iitp.ru

Received 7 February 2006
Revised 17 May 2006
Accepted 21 June 2006

Conservation rates in non-protein-coding regions of five yeast genomes of the genus *Saccharomyces* were analyzed using multiple whole-genome alignments. This analysis confirmed previously shown decrease in conservation rates observed immediately upstream of the translation start point and downstream of the stop-codon. Further, there was a sharp conservation peak in the upstream regions likely related to the core promoter (−35 bp to +35 bp around TSS) and a conservation peak downstream of the stop-codon whose function is not yet clear. Regulation of leucine and methionine biosynthesis controlled by the global regulator *Gcn4p* and pathway-specific regulators was analyzed in detail. A candidate alpha-isopropylmalate carrier, *YOR271cp*, was identified based on conservation of *Leu3p* binding sites, analysis of *ChIP*-chip data, protein localization and sequence similarity.

Keywords: Regulation of transcription; alpha-isopropylmalate transporter; evolution of regulatory signals.

1. Introduction

Extracting the complete functional information encoded in a genome — including genic, regulatory, and structural elements — is the central challenge in bioinformatics. Prediction of non-protein-coding functional regions, such as regulatory sites, is especially difficult because they are usually short (6-15 bp for *S. cerevisiae* and many other eukaryotic genomes), often degenerate, and can reside on either strand of DNA at variable distances from the genes they control. As functional sequences tend to be conserved through the evolution, they can appear as "phylogenetic footprints" in alignments of genome sequences of different species.[1] In that way transcription factor binding sites could be predicted to reside within such conserved footprints.

Shabalina *et al.*[2] observed conservation in 5′- and 3′- untranslated regions (UTR) at the large-scale level ascribing it to "common" functional sequences such as mRNA-ribosome interaction sites. Two groups used multiple whole-genome analysis[3] and multiple alignments of gene upstream regions[4] to identify binding signals for transcription factors. The results were represented as two lists of predicted binding motifs. The comparison of these lists shows rather moderate intersection. This prompted us to analyze the conservation rate for known and predicted binding sites in the *Saccharomyces* genomes in more detail.

In this paper, we describe the conservation rates of binding sites for transcriptional regulators of two metabolic pathways, biosynthesis of methionine and leucine. The amino acid biosynthetic pathways in yeast are subject to general transcriptional control by the global regulator of amino acid biosynthesis, *Gcn4p*. Its translation is activated in conditions of starvation for any of the ten amino acids.[5] The regulator binds to *Gcn4p*-responsive elements (GREs), with the core sequence TGACTC, in the upstream regions of regulated genes.[5–8] *Gcn4p* activates a variety of genes involved in amino acid biosynthesis.

In addition to *Gnc4p*, the amino acids biosynthesis pathways are controlled by pathway-specific regulators. The specific regulators of the methionine biosynthesis are the regulatory complex *Cbf1/Met4/Met28* and orthologous gene pair *Met31/Met32*. The binding site cores for these regulators are TCACGTG and AAACTGTGG, respectively.[9,10]

Leu3p is a specific regulator of the leucine biosynthesis. Its binding site consensus is CCG-N_4-CGG.[11] Several genes of this metabolic pathway (such as *Leu4, Ilv2, Ilv3*, and *Ilv5*) are controlled by both *Gcn4p* and *Leu3p*, whereas others are regulated only by one of these factors. For example, the expression of *Ilv1* does not depend on

Leu3p, and is under *Gcn4p* control only, while genes *Leu1* and *Leu2* are activated exclusively by *Leu3p*.[12]

De novo leucine biosynthesis includes four steps. The first pathway-specific reaction is catalyzed by alpha-isopropylmalate synthases, which are expressed in the absence of leucine and produce alpha-isopropylmalate.[13] The *Leu4* gene encodes two isoforms of alpha-isopropylmalate synthase due to alternative transcription start sites. The longer isoform is localized in the mitochondrial matrix and the shorter one, in the cytoplasm.[14] Both isoforms are functional and could produce alpha-isopropylmalate.[15] However, in yeast, the enzymes involved in the *de novo* synthesis of alpha-isopropylmalate are largely associated with the mitochondria,[16] and the role of the cytosolic isoform of alpha-isopropylmalate synthase in not yet clear. It may be needed only under anaerobic conditions, when the mitochondrial isoform might be unstable or nonfunctional.[17]

Thus, alpha-isopropylmalate (at least partially) is synthesized in the mitochondrial matrix. Subsequent stages of the leucine biosynthesis occur in the cytoplasm,[18] so this intermediate has to be transported into the cytosole. Moreover, alpha-isopropylmalate is a co-activator of regulator *Leu3p*,[13,19–21] and thus should be transferred into the nucleus as well. To our knowledge, no alpha-isopropylmalate transporter of yeast has been experimentally identified. Using the comparative analysis of regulation and other supporting evidence, we have identified a candidate gene for the alpha-isopropylmalate carrier.

2. Material and Methods

2.1. *Global analysis*

We used the yeast genome annotation extracted from the SGD database (ftp://genome-ftp.stanford.edu/pub/yeast/data_download/chromosomal_feature/saccharomyces_cerevisiae.gff) to map 6578 ORFs to the finished genome of *S. cerevisiae*.[22] The draft genomes of four species of the *Saccharomyces* sensu-stricto group (*S. paradoxus, S. mikatae, S. bayanus*,[3] and *S. kudriavtsevii*[4]) were aligned to the genome of *S. cerevisiae* using the MLAGAN program[23] as described in Ref. 24. In the analysis of conservation of the 5′ upstream (3′ downstream) regions of yeast genes, we considered only regions not overlapping with other genes. The maximum length of the analyzed regions was 500 nucleotides. The alignments of the upstream and downstream regions for all considered genes are available at http://www.rtcb.iitp.ru/kovaleva_e.htm.

We used the SCPD database[25] to map 452 binding sites for 90 transcription factors in the upstream regions of 166 genes of *S. cerevisiae*. For each gene, we excluded binding sites that mapped outside the considered upstream region.

For position i in the alignment, conservation C_i was defined as

$$C_i = \frac{m_i}{N - g_i}$$

where m_i and g_i are the numbers of matches to *S. cerevisiae* and gaps in the alignment column, respectively, and N is the total number of non *cerevisiae* species in the alignment.

2.2. *Analysis of individual regulatory systems*

Fragments covering 750 nucleotides of upstream regions and 150 nucleotides of protein-coding genes of *S. cerevisiae* were considered. Search for orthologs was done using fungiBLAST (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi?organism=fungi). Protein-coding genes without identifiable orthologs were ignored. Conversely, regions upstream of orthologous genes were considered even if they did not produce a strong alignment.

Genome Explorer[26] was used to identify candidate sites. SignalX[27] was used to construct positional weight matrices. As our matrices are similar to the already published ones,[28] we did not perform detailed comparison of the matrix specificity. New matrices were constructed to facilitate the work with our tools. Multiple sequence alignments were done using ClustalX.[29] Binding sites for transcription factors of *S. cerevisiae* were taken from the TRANSFAC[30] (http://www.gene-regulation.com/pub/databases.html#transfac) and SCPD[25] (http://rulai.cshl.edu/SCPD/) databases. Transmembrane domains were predicted using TMHMM[31] (http://www.cbs.dtu.dk/services/TMHMM/) and PSORT[32] (http://psort.nibb.ac.jp/). Phylogenetic tree was constructed by the maximum likelihood method implemented in PHYLIP.[33]

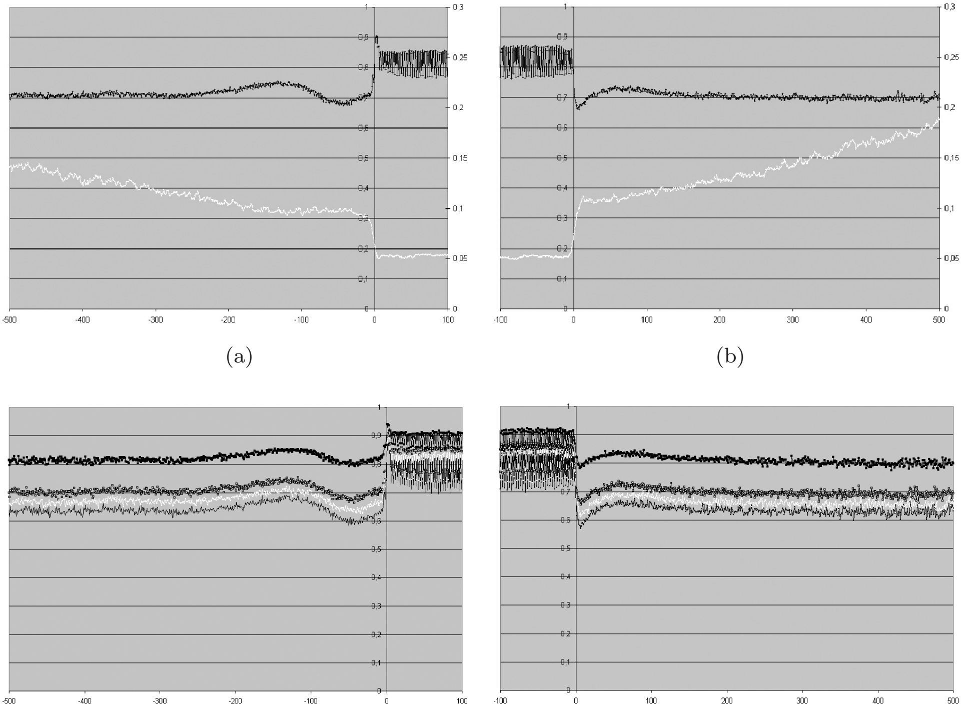
3. Results and Discussion

3.1. *Conservation rates in untranslated regions*

Using the MLAGAN program, we constructed multiple whole-genome alignment of five *Saccharomyces* genomes: *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. bayanus,* and *S. kudriavzevii* (see Material and Methods). Multiple alignments for *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, and *S. bayanus* intergenic regions have been published earlier.[3] The results obtained for this alignment and the MLAGAN aligment are similar (data not shown). The results reported below were obtained using the MLAGAN alignment.

We first analyzed the positional conservation rates in 5′. We confirmed relatively high-conservation rate in the region immediately upstream of the start codon observed in Ref. 2. In addition to this conservation peak, we observed a region of higher conservation further upstream, starting at about 70 bp, reaching the maximum at 120 bp and getting back to the background level at 220 bp upstream of the translation start point (Figs. 1(a) and 1(b)). This conservation peak could not be explained by accumulation of transcription factor binding sites, as the distribution of these sites reaches maximum at 200 bp upstream of the translation start point and returns to the background level beyond 600 bp upstream of start codon (data not shown). We suppose that the observed conservation at the (−220–70 bp) region may be caused by binding sites for TATA-box-binding proteins (core promoters). The distribution of these sites is similar to the conservation plot in this region (data not shown).

As for downstream intergenic regions, our results extend the observations by Shabalina *et al.*[2] They described a decrease in the conservation rate immediately downstream of the stop codon. We analyzed the conservation rates in larger untranslated regions. Our analysis confirmed the decrease in the conservation level observed by Shabalina *et al.*,[2] but we also observed an increase in the conservation rate covering the region (25–155 bp) downstream of the coding region (Figs. 1(c) and (d)). Conservation is gradually increasing in the interval from 25 bp downstream of the stop codon to 50 bp downstream of the stop codon, then flattens out and starts to



(a)

(b)

(c)

(d)

Fig. 1. Positional conservation rates in untranslated regions. (a) and (b). Average conservation rates in 5′-UTRs and 3′-UTRs, respectively, of all five Saccharomyces genomes. Black — the conservation rates per nucleotide (central axis), white — the density of gaps (right axis). (c) and (d). Conservation rates in 5′-UTR and 3′-UTR, respectively, of four *Saccharomyces* genomes compared to *S. cerevisiae* individually. Black squares — *S. paradoxus* genome, empty squares — *S. mikatae*, white curve — *S. kudriavzevii*, black curve — *S. bayanus.* Zero in (a) and (c) corresponds to the translational start point and positive numbers corresponds to nucleotide positions in the coding region. Zero in (b) and (d) corresponds to the stop codon and negative numbers correspond to the nucleotide positions in coding region.

decrease gradually around 100 bp downstream of the stop codon until reaching the background level at about 155 bp downstream of the stop codon.

Thus, in fact, the region of decreased conservation is followed by a long-tail peak in the conservation rate in the 3′-UTR. It is likely that peak is caused by mRNA stability and localization sites that are concentrated in the region 50–100 nt downstream of the stop-codon.[34,35]

3.2. *Conservation of transcription factor-binding sites*

We studied the positional conservation rates in alignments centered at known binding sites for transcription regulators present in the SCPD database. This database includes experimentally mapped binding sites of length varying from five to 56 nucleotides. As expected, the conservation rate at transcription factors binding sites was much higher than the conservation rate in surrounding non regulatory parts of upstream regions. The conservation rate was highest in the middle of the sites and decreasing gradually toward the surrounding region. However, it remained elevated in the vicinity of the sites, extending to approximately 100 bp (Fig. 2). This region of elevated conservation around the sites cannot be attributed to imprecise positioning of sites in DNA footprinting experiments, since the region of increased conservation extended beyond the length of the longest sites (Fig. 2). It is also unlikely to be an artifact linked with anchoring of alignments by highly conserved binding sites,[2]

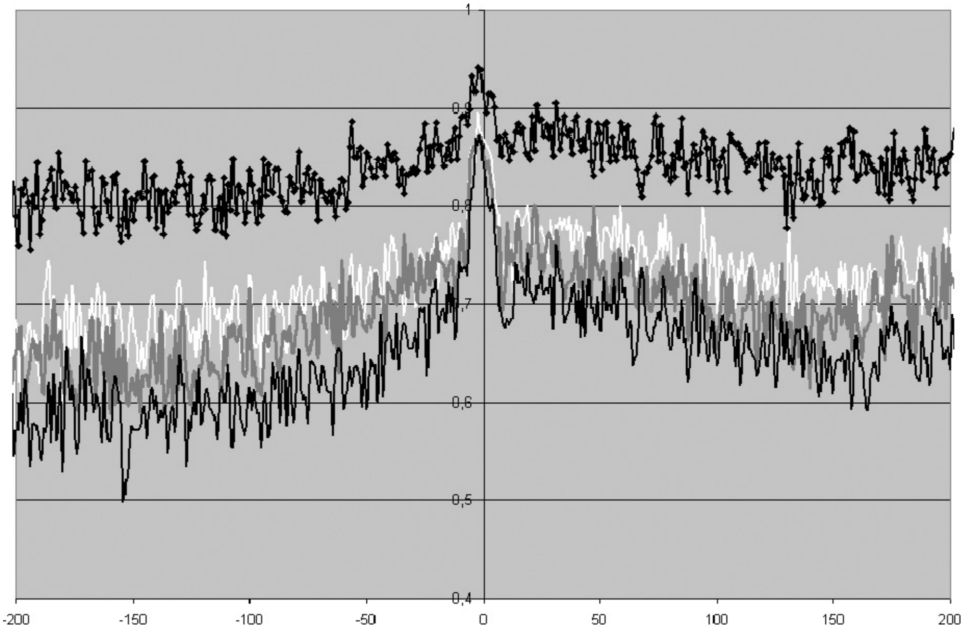


Fig. 2. Positional conservation rates for alignment centered at binding sites of varying length in each genome separately. Zero corresponds to the site center. Black squares — *S. paradoxus*, white curve — *S. mikatae*, grey curve — *S. kudriavzevii*, and black curve — *S. bayanus*.

since we excluded positions containing gaps from analysis. The observed elevated conservation level at 5′ of the site also could not be explained by the general decrease in the conservation rate caused by extended distance from the gene, as our results suggest that increasing the upstream distance does not lead to a noticeable decrease in the conservation rate. Thus, we conclude that observed elevated conservation 5′ may be due to the presence of other binding sites or promoters. Consistent with this explanation, the conservation rate in the region between the site and the gene (3′ of the site) is higher than in the region further upstream (5′ of the site). As binding sites for transcription regulators in eukaryotes are usually short, clustering of transcription factor binding sites is one of the mechanisms providing the specific regulation via cooperative binding of transcription factors to DNA strands.

3.3. *Conserved regulation of the methionine and leucine biosynthesis pathways*

Our observation of the extended region of higher conservation around the binding sites for transcription factors could, at least partially, explain the differences in the sets of predicted signals in Refs. 3 and 4, as different algorithms may pick up different conserved sites in these regions. Another possible explanation could be that the binding sites are not absolutely conserved, and again, their determination is algorithm-dependent. To examine this possibility, we analyzed in detail the conservation of binding sites for transcriptional activators of two amino acids biosynthesis pathways, biosynthesis of methionine and leucine.

Experimentally verified binding sites for transcription regulators are unknown for *Saccharomyces* genomes other than *S. cerevisiae,* and we could not construct position weight matrices for these genomes. Multiple protein alignments of DNA-binding regions of regulators indicate high similarity of the binding region in all species (data not shown). Thus, we believe that the binding signals are very similar in all studied genomes, and the positional weight matrices constructed for binding sites of *S. cerevisiae* could be applied to other genomes as well. We investigated the conservation rates of binding sites in multiple alignments of upstream regions.

We assumed that a binding site is conserved if it was aligned to a candidate site in the other genomes. In order to take into account the possibility of large insertions in the upstream regions leading to misalignment, we independently searched for candidate sites in all genomes using the *S. cerevisiae* matrices and then aligned the regions around the candidate sites. However, all such sites were observed in dissimilar, likely non homologous regions, and thus they seem to be false positives.

To simplify the estimations, we used the quantitative assessment of conservation rates. If a binding site was exactly conserved in a studied genome (compared to *S. cerevisiae*) we scored it as 1. If a site is not exactly conserved, but still aligns to a candidate site recognized by the respective profile it was scored as 0.5. We tabulated the sums of the total site scores for transcription regulators separately for each genome.

The conservation rates of binding sites for the global regulator of amino acid biosynthesis, *Gcn4p*, were examined in regulatory regions of nine genes with experimentally verified binding sites listed in the SCPD database: *HIS3*, *ARG8*, *ARG1*, *ADE4*, *ILV1*, *TRP4*, *HIS4*, *HIS7*, and *ILV2*. Known binding sites upstream of these genes were used to construct the positional weight matrix.

As mentioned above, the biosynthesis of methionine is regulated by three more or less independent regulators or regulatory complexes: *Gcn4p, Met31/Met32*, and *Cbf1/Met4/Met28*. Matrices for the *Cbf1/Met4/Met28* and *Met31/Met32* binding sites were constructed using sites from the SCPD database and the known consensus, respectively. Then these matrices were applied to genes known to be regulated by these factors.[9]

Similarly, structural genes of leucine biosynthesis are independently regulated by *Gcn4p* and *Leu3p*, and these genes could be activated by both or any one of these regulators. So, candidate binding sites for these transcription factors were predicted only in the upstream regions of genes known to be regulated by a particular regulator. The *Leu3p* binding sites are not listed in the TRANSFAC database, and we constructed a position weight matrix on the basis of the consensus[36] and several experimentally verified binding sites.[11,12,14,37–39]

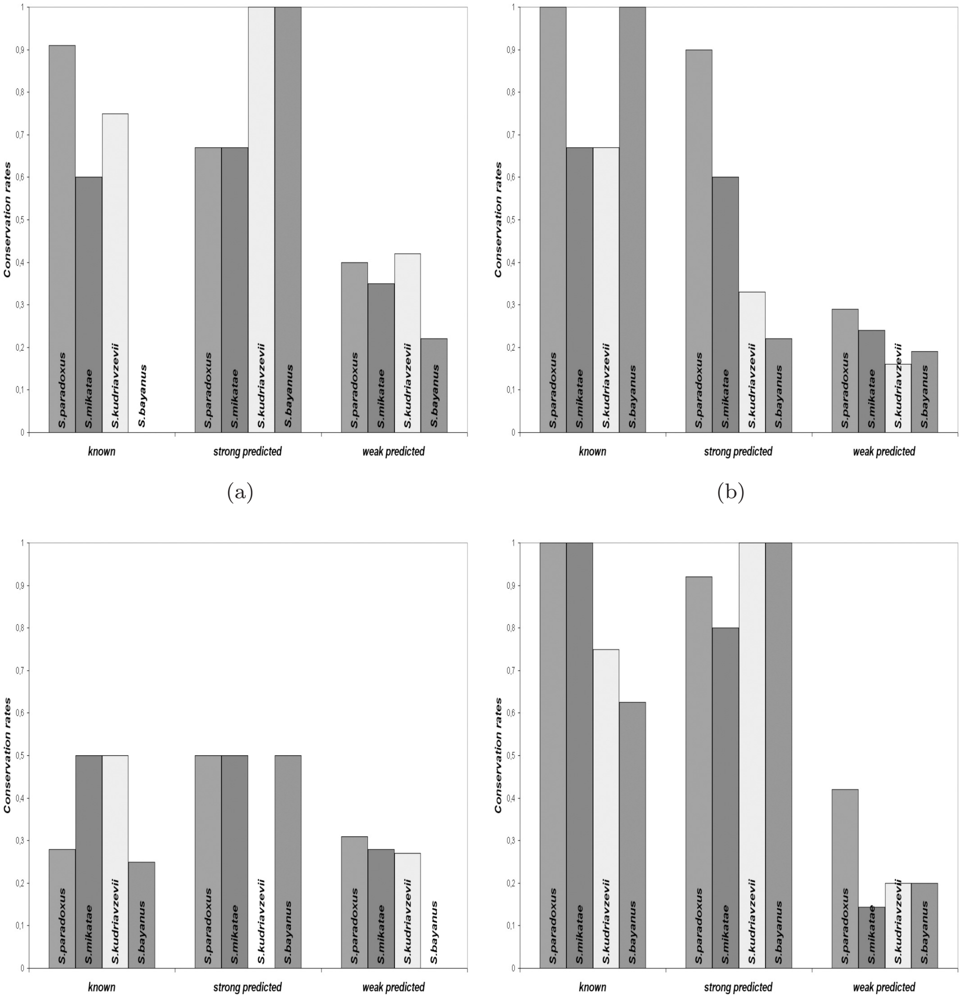
All identified sites were divided into three groups: "known" (experimentally verified sites), "strong predicted sites" (candidate sites with a score exceeding that of the majority of known sites), and "weak predicted" sites (candidate sites with a lower score, but still higher than the threshold set to the minimal score of experimentally verified sites).

As the considered genomes are not completely sequenced, we normalized the conservation rate dividing the total score by the number of orthologous genes for which the upstream regions were available. For genes having multiple sites for one regulator, the conservation of at least one site was considered to be sufficient for conservation of the regulatory interaction.

The average conservation rates for all studied regulators are presented in Fig. 3. Our data demonstrate that, the conservation rates of known and strong predicted binding sites are similar for each regulator (except the *Met31/Met32*, see below, and some rare cases when no orthologs could be identified). Still, even strongest sites are not necessarily conserved in all examined genomes, as it has been implicitly assumed in the cited studies. Weak predicted binding sites serve as a control reflecting conservation rates for "random" sites.

As for the *Met31/Met32* regulator(s), the conservation rates for all groups of sites seem to be rather low even in such closely related genome as *S. paradoxus.* We could not reliably explain such sharp decrease in the conservation of binding sites for this regulatory complex, as even the molecular function of this regulatory complex is still unclear.[9,40]

On the other hand, the conservation rates of known and predicted binding sites of *Leu3p* are much higher than the average conservation rate of binding sites for other regulators. This difference can be due to the length of the signal: for *Leu3* it



(a)

(b)

(c)

(d)

Fig. 3. Average conservation rates of binding sites for all four studied regulators. Conservation rates are shown individually for each group of sites (such as "weak", etc.) for each regulator. (a) *Gcn4p* binding sites; (b) *Cbf1*-complex sites; (c) Binding sites of the *Met31/Met32*; (d) *Leu3p* binding sites. See the text for definitions.

seems to be about ten nucleotides, as the internal positions also contribute to the strength of the interaction.

But other explanations are also possible. For instance, in most cases there are multiple copies of *Gcn4p* binding sites in the upstream regions of controlled genes. This could increase the overall size and specificity of the effective binding.[6] The complete disruption or reduction of affinity of some of these sites would not make a gene misregulated. On the contrary, there is the only one site for *Leu3p* in an upstream region of each regulated gene, and thus the complete or partial destruction of this site should abolish the transcriptional regulation of the gene. This situation reflects that functional constraints on the "important" sites, such as the sites for *Leu3p*, are much stronger than on each of the multiple sites for *Gcn4p*. Therefore, this could explain the difference between the conservation rates of the binding sites for these particular regulators.

Thus, our detailed analysis shows that binding sites for eukaryotic transcription regulators may not be entirely conserved, and the standard phylogenetic footprinting techniques cannot be applied to yeast genomes without correction. Although the conservation rates of known and strong predicted binding sites are similar, even the strongest sites (including experimentally verified ones) may not be conserved even in the closest genome.

3.4. *Identification of a candidate alpha-isopropylmalate transporter*

We used the constructed positional weight matrices to identify other genes that could be involved in the methionine and leucine biosynthesis. This allowed us to identify *YOR271c* as the gene that could encode the transporter for alpha-isopropylmalate, an intermediate of the leucine biosynthesis. This prediction is based on the following observations:

- The protein product of this gene is known to localize in the mitochondrion.[41]
- *YOR271cp* contains several predicted transmembrane domains (Fig. 4) and thus is structurally likely to be a transporter.
- The predicted binding site for the leucine pathway transcriptional regulator, *Leu3p*, is strongly conserved in all studied genomes and is identical to the binding site consensus (Fig. 5).
- The ChIP microarray data shows that *YOR271c* is activated by *Leu3p* along with several other structural genes involved in the leucine biosynthesis.[42] We

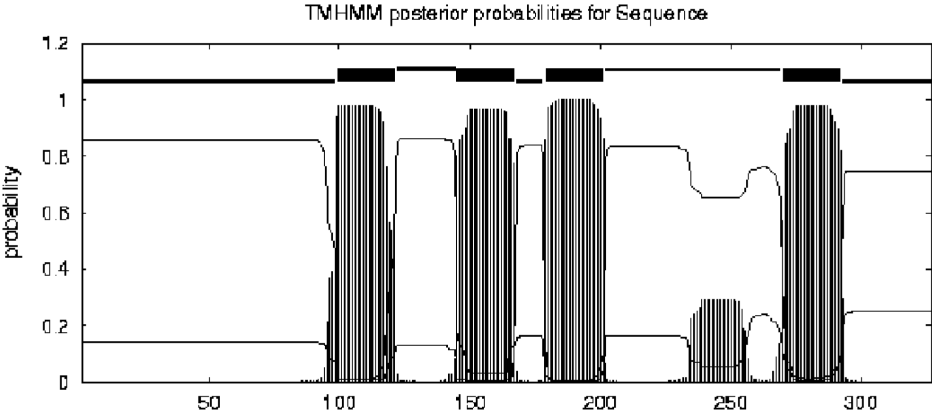


Fig. 4. Transmembrane segments predicted for *YOR271cp*.

```
Scer_YOR271c____TC-GTACTGGCGTCCGGTACCGGAGCGT-----ACCA---GA
Spar_YOL271c____CT-GTACTGGCGTCCGGTACCGGACCGT-----ACCA---GC
Skudr_YOR271c___CCAGCACCGTCTTCCGGTACCGGAGCGTC-GCTTCCC---A-
Sbay_YOR271c____TTGGCGCCGTCTACCGGTACCGGACCACT-GCGGCCA---GC
Smik_YOR271c____CCGGCACTGGCCTCCGGTACCGGACCGTCCGTTGCCCAGCGC
                   *  * * *  *********** *        **
```

Fig. 5. Alignment of upstream regions of the *YOR271c* gene of *S. cerevisiae* and its orthologs from the *S. paradoxus*, *S. kudriavzevii*, *S. bayanus*, and *S. mikatae* genomes. Predicted binding sites for *Leu3p* are colored in grey. Conserved positions are marked by asterisks.

Table 1. Analysis of expression of the *Leu3p*-regulated genes in the *ChIP* microarray data from Ref. 42. (a) Genes activated by *Leu3p* (p-value < 0.001). Genes involved in the leucine biosynthesis pathway[17] are colored in grey, and the predicted transporter for alpha-isopropylmalate is shown in bold. (b) Genes known to be regulated by *Leu3p*, which did not respond to the *Leu3p* expression.

Molecular functions of the genes were obtained from *Saccharomyces* genome database (http://www.yeastgenome.org).

Gene Name	ORF	Molecular Function
		(a)
OAC1	YKL120W	Mitochondrial inner membrane transporter, transports oxaloacetate, sulfate, and thiosulfate
LEU1	YGL009C	Isopropylmalate isomerase, catalyzes the second step in the leucine biosynthesis pathway
SET5	YHR207C	Molecular function unknown
BAT1	YHR208W	Mitochondrial branched-chain amino acid aminotransferase
ILV2	YMR108W	Acetolactate synthase, localizes in the mitochondria; expression of the gene is under general amino acid control
YKL118W	YKL118W	Molecular function unknown
YDR042C	YDR042C	Molecular function unknown
BAP2	YBR068C	High-affinity leucine permease, functions as a branched-chain amino acid permease involved in the uptake of leucine, isoleucine, and valine
YOR271C	**YOR271C**	**Molecular function unknown**
MET4	YNL103W	Lecine-zipper transcriptional activator, responsible for the regulation of the sulfur amino acid pathway
LEU4	YNL104C	Alpha-isopropylmalate synthase (2-isopropylmalate synthase); the main isozyme responsible for the first step in the leucine biosynthesis pathway
SPS2	YDR522C	Molecular function unknown
LEU9	YOR108w	Alpha-isopropylmalate synthase II (2-isopropylmalate synthase), catalyzes the first step in the leucine biosynthesis pathway

Table 1. (*Continued*)

Gene Name	ORF	Molecular Function
YDR210W-D	YDR210W-D	TyB Gag-Pol protein; proteolytically processed to make the Gag, RT, PR, and IN proteins that are required for retrotransposition
RRP6	YOR001W	Exonuclease component of the nuclear exosome; contributes to the quality-control system that retains and degrades aberrant mRNAs in the nucleus
RML2	YEL050C	Mitochondrial ribosomal protein of the large subunit, has similarity to *E. coli* L2 ribosomal protein
MRPL24	YMR193W	Structural constituent of ribosome
YNL050C	YNL050C	Molecular function unknown
YDL228C	YDL228C	Molecular function unknown
YHR209W	YHR209W	Putative *S*-adenosylmethionine-dependent methyltransferase of the seven beta-strand family
YCR018C-A	YCR018C-A	Molecular function unknown
MAK32	YCR019W	Molecular function unknown
		(b)
LEU2	YCL018w	Beta-isopropylmalate dehydrogenase, catalyzes the third step in the leucine biosynthesis pathway
ILV6	YCL009C	Regulatory subunit of acetolactate synthase, which catalyzes the first step of branched-chain amino acid biosynthesis
ILV3	YJR016C	Dihydroxyacid dehydratase, catalyzes third step in the common pathway leading to biosynthesis of branched-chain amino acids
ILV5	YLR355c	Acetohydroxyacid reductoisomerase, mitochondrial protein involved in branched-chain amino acid biosynthesis
GDH1	YOR375c	NADP(+)-dependent glutamate dehydrogenase, synthesizes glutamate from ammonia and alpha-ketoglutarate; rate of alpha-ketoglutarate utilization differs from Gdh3p; expression regulated by nitrogen and carbon sources
BAT2	YJR148W	Cytosolic branched-chain amino acid aminotransferase
GAP1	YKR039w	General amino acid permease
MAE1	YKL029c	Mitochondrial malic enzyme, catalyzes the oxidative decarboxylation of malate to pyruvate, which is a key intermediate in sugar metabolism and a precursor for synthesis of several amino acids

analyzed these data in more detail to ensure that there are no other candidates for the role of alpha-isopropylmalate transporter (Table 1). About half of the genes identified in the experiment are known genes of the leucine pathway (Table 1a), whereas about half of the genes known to be regulated by *Leu3p* were missed (Table 1b).

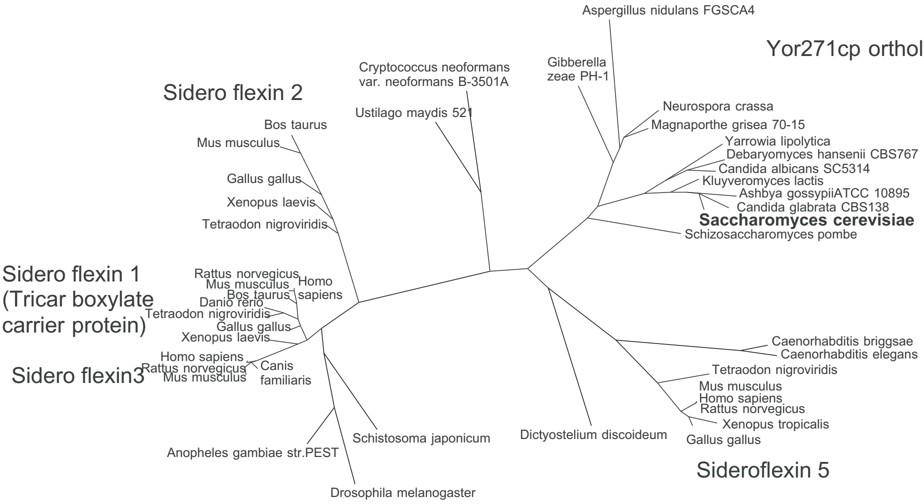


Fig. 6. Phylogenetic tree of YOR271cp homologs. YOR271cp is shown in bold.

Searching for possible alternative candidates, we considered *ChIP*-identified genes with unknown molecular function. We applied the *Leu3p* positional weight matrix to these nine genes. Only two of them, *YOR271c* (our candidate) and *YDL228c*, of *S. cerevisiae* had potential binding sites for *Leu3p*. The site upstream of *YDL228c* was weaker and was not conserved even in the closest genomes.

Thus, *YOR271c* might be a transporter involved in the leucine biosynthesis.

Further, we constructed a phylogenetic tree of *YOR271c* homologs (Fig. 6). The closest *YOR271c* homologs with known function are tricarboxylate transporter from *Rattus norvegicus* and sideroflexins.

The published data on sideroflexins are conflicting. The mouse gene encoding *sideroflexin-1 (Sfxn1)* was discovered during investigation of mutation that induces syderocytic anemia,[43] a disorder associated with aberrant mitochondrial iron homeostasis.[44] Although Fleming *et al.*[43] mentioned high similarity of the sideroflexin with the rat tricarboxylate transporter, they assumed that the gene they had cloned did not encode a tricarboxylate transporter, since a constitutively expressed mitochondrial tricarboxylate carrier was known.[45] They further suggested that the rodent sideroflexin homolog had been ascribed a wrong function of a tricarboxylate transporter as a consequence of co-extraction of this protein with the constitutive tricarboxylate carrier. However, during extraction of the presumed rat transporter for tricarboxylates, it had been analyzed for the substrate specificity and its ability to transfer tricarboxylates had been confirmed.[46]

Overall, although each observation is rather weak, we believe that taken together they are sufficiently convincing to warrant experimental verification.

4. Conclusions

In summary, our analysis of conservation of upstream regions of the genes and of individual transcription factor binding sites reveals a pattern of elevated conservation of regions of non coding DNA with regulatory function, compared to background conservation level. This conservation is not limited to individual binding sites, as shown previously most notably in studies,[47–49] but extends much further into the surrounding region. As expected, the conservation of all regions decreases with increased phylogenetic distance between the analyzed species. The phylogenetic footprint of the binding sites is most obvious at intermediate phylogenetic distances. Regions of increased conservation of individual binding sites extend beyond the edges of these sites, possibly indicating distributed clustering of the binding sites.

Pattern of conservation of individual sites shows relative decrease of conservation in distant genomes and variability of conservation rates of sites for different transcription factors. At that, longer single sites are more conserved than short, clustered sites. Recently, a considerable evolvability has been demonstrated for the regulation of mitochondrial and cytoplasmic ribosomal proteins.[50,51] Our study demonstrates that the same seem to hold for primary metabolic pathways. Functional relevance of non consensus sites was demonstrated for transcription factors involved in the regulation of cell cycle and development.[52]

Nevertheless, analysis of regulatory patterns combined with other large-scale data allowed for identification of a candidate mitochondrial transporter for alpha-isopropylmalate.

Acknowledgments

We are grateful to Renata Zvyagilskaya for useful discussions and to Mukund Sundarajan for help with the alignments. This study was partially supported by grants from the Howard Hughes Medical Institute, Ludwig Institute for Cancer Research, the Fund for Support of Russian Science, and the Russian Academy of Sciences (program "Molecular and Cellular Biology"). G.B. gratefully acknowledges fellowships from the Pew Charitable Trusts award 2000-002558 and the Burroughs Wellcome Fund award 1001782, both to Princeton University.

References

1. Hardison RC, Oeltjen J, Miller W, Long human-mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome, *Genome Res* **7**:959–966, 1997.
2. Shabalina SA, Ogurtsov AY, Rogozin IB, Koonin EV, Lipman DJ, Comparative analysis of orthologous eukaryotic mRNAs: Potential hidden functional signals, *Nucleic Acids Res* **32**:1774–1782, 2004.
3. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES, Sequencing and comparison of yeast species to identify genes and regulatory elements, *Nature* **423**:241–254, 2003.

4. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M, Finding functional features in Saccharomyces genomes by phylogenetic footprinting, *Science* **301**:71–76, 2003.
5. Hinnebusch AG, Natarajan K, Gcn4p, a master regulator of gene expression, is controlled at multiple levels by diverse signals of starvation and stress, *Eukaryot Cell* **1**:22–32, 2002.
6. Arndt K, Fink GR, GCN4 protein, a positive transcription factor in yeast, binds general control promoters at all 5′ TGACTC 3′ sequences, *Proc Natl Acad Sci USA* **83**:8516–8520, 1986.
7. Hinnebusch AG, Mechanisms of gene regulation in the general control of amino acid biosynthesis in Saccharomyces cerevisiae, *Microbiol Rev* **52**:248–273, 1988.
8. Natarajan K, Meyer MR, Jackson BM, Slade D, Roberts C, Hinnebusch AG, Marton MJ, Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast, *Mol Cell Biol* **21**:4347–4368, 2001.
9. Thomas D, Surdin-Kerjan Y, Metabolism of sulfur amino acids in Saccharomyces cerevisiae, *Microbiol Mol Biol Rev* **61**:503–532, 1997.
10. Gonze D, Pinloche S, Gascuel O, van Helden J, Discrimination of yeast genes involved in methionine and phosphate metabolism on the basis of upstream motifs, *Bioinformatics* **21**:3490–3500, 2005.
11. Hellauer K, Rochon MH, Turcotte B, A novel DNA binding motif for yeast zinc cluster proteins: The Leu3p and Pdr3p transcriptional activators recognize everted repeats, *Mol Cell Biol* **16**:6096–6102, 1996.
12. Tu H, Casadaban MJ, The upstream activating sequence for L-leucine gene regulation in *Saccharomyces cerevisiae*, *Nucleic Acids Res* **18**:3923–3931, 1990.
13. Chang LF, Cunningham TS, Gatzek PR, Chen WJ, Kohlhaw GB, Cloning and characterization of yeast Leu4, one of two genes responsible for alpha-isopropylmalate synthesis, *Genetics* **108**:91–106, 1984.
14. Beltzer JP, Chang LF, Hinkkanen AE, Kohlhaw GB, Structure of yeast LEU4. The 5′ flanking region contains features that predict two modes of control and two productive translation starts, *J Biol Chem* **261**:5160–5167, 1986.
15. Beltzer JP, Morris SR, Kohlhaw GB, Yeast LEU4 encodes mitochondrial and non-mitochondrial forms of alpha-isopropylmalate synthase, *J Biol Chem* **263**:368–374, 1988.
16. Ryan ED, Kohlhaw GB, Subcellular localization of isoleucine-valine biosynthetic enzymes in yeast, *J Bacteriol* **120**:631–637, 1974.
17. Kohlhaw GB, Leucine biosynthesis in fungi: Entering metabolism through the back door, *Microbiol Mol Biol Rev* **67**:1–15, 2003.
18. Ryan ED, Tracy JW, Kohlhaw GB, Subcellular localization of the leucine biosynthetic enzymes in yeast, *J Bacteriol* **116**:222–225, 1973.
19. Hu Y, Kohlhaw GB, Additive activation of yeast LEU4 transcription by multiple cis elements, *J Biol Chem* **270**:5270–5275, 1995.
20. Hu Y, Cooper TG, Kohlhaw GB, The *Saccharomyces cerevisiae Leu3* protein activates expression of *GDH1*, a key gene in nitrogen assimilation, *Mol Cell Biol* **15**:52–57, 1995.
21. Wang D, Hu Y, Zheng F, Zhou K, Kohlhaw GB, Evidence that intramolecular interactions are involved in masking the activation domain of transcriptional activator *Leu3p*, *J Biol Chem* **272**:19383–19392, 1997.
22. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG, Life with 6000 genes, *Science* **274**:563–567, 1996.

23. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S, LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA, *Genome Res* **13**:721–31, 2003.
24. Sundararajan M, Brudno M, Small K, Sidow A, Batzoglow S, *Proceedings of the fourth Workshop on Algorithms in Bioinformatics (WABI 2004)*, 2004.
25. Zhu J, Zhang MQ, SCPD: A promoter database of yeast *Saccharomyces cerevisiae*, *Bioinformatics* **15**:607–611, 1999.
26. Mironov AA, Koonin EV, Roytberg MA, Gelfand MS, Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes, *Nucleic Acids Res* **27**:2981–2989, 1999.
27. Mironov AA, Vinokurova NP, Gelfand MS, GenomeExplorer: Software for analysis of complete bacterial genomes, *Mol Biol* **34**:222–231, 2000.
28. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA, Transcriptional regulatory code of a eukaryotic genome, *Nature* **431**:99–104, 2004.
29. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG, The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Res* **25**:4876–4882, 1997.
30. Wingender E, Kel AE, Kel OV, Karas H, Heinemeyer T, Dietze P, Knüppel R, Romaschenko AG and Kolchanov NA, TRANSFAC, TRRD and COMPEL: Towards a federated database system on transcriptional regulation, *Nucleic Acids Res* **25**:265–268, 1997.
31. Krogh A, Larsson B, von Heijne G, Sonnhammer EL, Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes, *J Mol Biol* **305**:567–580, 2001.
32. Horton P, Nakai K, A probabilistic classification system for predicting the cellular localization sites of proteins, *Intellig Syst Mol Biol* **4**:109–115, 1996.
33. Felsenstein J, Evolutionary trees from DNA sequences: A maximum likelihood approach, *J Mol Evol* **17**:368–376, 1981.
34. Graber JH, Variations in yeasts 3′-processing cis-elements correlate with transcript stability, *Trends Genet* **19**:473–476, 2003.
35. Shalgi R, Lapidot M, Shamir R, Pilpel Y, A catalog of stability-associated sequence elements in 3′ UTRs of yeast mRNAs, *Genome Biol* **6**:R86, 2005.
36. Zhou K, Brisco PR, Hinkkanen AE, Kohlhaw GB, Structure of yeast regulatory gene LEU3 and evidence that LEU3 itself is under general amino acid control, *Nucleic Acids Res* **15**:5261–5273, 1987.
37. Falco SC, Dumas KS, Livak KJ, Nucleotide sequence of the yeast ILV2 gene which encodes acetolactate synthase, *Nucleic Acids Res* **13**:4011–4027, 1985.
38. Hsu YP, Schimmel P, Yeast LEU1. Repression of mRNA levels by leucine and relationship of 5′-noncoding region to that of LEU2, *J Biol Chem* **259**:3714–3719, 1984.
39. Petersen JG, Holmberg S, The ILV5 gene of Saccharomyces cerevisiae is highly expressed, *Nucleic Acids Res* **14**:9631–9651, 1986.
40. Blaiseau PL, Isnard AD, Surdin-Kerjan Y, Thomas D, *Met31p* and *Met32p*, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism, *Mol Cell Biol* **17**:3640–3648, 1997.
41. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK, Global analysis of protein localization in budding yeast, *Nature* **425**:686–691, 2003.
42. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon

DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA, Transcriptional regulatory networks in Saccharomyces cerevisiae, *Science* **298**:799–804, 2002.

43. Fleming MD, Campagna DR, Haslett JN, Trenor III CC, Andrews NC, A mutation in a mitochondrial transmembrane protein is responsible for the pleiotropic hematological and skeletal phenotype of *flexed-tail (f/f)* mice, *Genes Dev* **15**:652–657, 2001.
44. Roy CN, Andrews NC, Recent advances in disorders of iron metabolism: Mutations, mechanisms and modifiers, *Hum Mol Genet* **10**:2181–2186, 2001.
45. Xu Y, Mayor JA, Gremse D, Wood DO, Kaplan RS, High-yield bacterial expression, purification, and functional reconstitution of the tricarboxylate transport protein from rat liver mitochondria, *Biochem Biophys Res Commun* **207**:783–789, 1995.
46. Azzi A, Glerum M, Koller R, Mertens W, Spycher S, The mitochondrial tricarboxylate carrier, *J Bioenerg Biomembr* **25**:515–524, 1993.
47. Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB, Position specific variation in the rate of evolution in transcription factor. binding sites, *BMC Evol Biol* **3**:19, 2003.
48. Gasch AP, Moses AM, Chiang DY, Fraser HB, Berardini M, Eisen MB, Conservation and evolution of cis-regulatory systems in ascomycete fungi, *PLoS Biol* **2**:e398, 2004.
49. Liu Y, Liu XS, Wei L, Altman RB, Batzoglou S, Eukaryotic regulatory element conservation analysis and identification using comparative genomics, *Genome Res* **14**:451–458, 2004.
50. Tanay A, Regev A, Shamir R, Conservation and evolvability in regulatory networks: The evolution of ribosomal regulation in yeast, *Proc Natl Acad Sci USA* **102**:7203–7208, 2005.
51. Ihmels J, Bergmann S, Gerami-Nejad M, Yanai I, McClellan M, Berman J, Barkai N, Rewiring of the yeast transcriptional network through the evolution of motif usage, *Science* **309**:938–940, 2005.
52. Doniger SW, Huh J, Fay JC, Identification of functional transcription factor binding sites using closely related Saccharomyces species, *Genome Res* **15**:701–709, 2005.

Galina Yu. Kovaleva is a graduate student in the Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia. She obtained her M.S. in Biochemistry from the Department of Virology of Moscow State University in 2003. Her research interests are in comparative genomics, molecular biology and molecular evolution.

Georgii A. Bazykin received his diploma (B.S.) in Biology from Moscow State University, Moscow, Russia in 2001. He is currently pursuing his Ph.D. degree at Princeton University. His interests are in molecular evolution, focusing on methods for inferring positive selection from sequence evolution.

Michael Brudno is an Assistant Professor and a Canada Research Chair in Computational Biology at the University of Toronto, holding a joint appointment in the Department of Computer Science and the Banting and Best Department of Medical Research. His main research interests are in the development of algorithms for biological data, especially biological sequences. Before coming to Toronto, he did his Postdoctoral Research and Undergraduate work at University of California at Berkeley, and his Ph.D. at Stanford University.

Mikhail S. Gelfand is the Head of the Research and Training Center in Bioinformatics of the Institute for Information Transmission Problems, RAS in Moscow, Russia, and a Professor at the Department of Bioengineering and Bioinformatics of the Moscow State University. He graduated from the Department of Mathematics of the Moscow State University, received his Ph.D. (Math.) degree from the Institute of Theoretical and Experimental Biophysics, RAS (Pushchino), and the Doctor of Sciences degree from the State Research Institute for Genetics and Selection of Industrial Microorganisms (Moscow). He is a member of editorial boards of several journals, in particular, “PLoS Biology”, “Bioinformatics”, “BMC Bioinformatics”, “Journal of Bioinformatics and Computational Biology”, and “Journal of Computational Biology”. He received the A. A. Baev prize (1999) from the Russian State “Human Genome” Council, and “The Best Scientist of the Russian Academy of Sciences” award (2004). His research interests include comparative genomics, metabolic reconstruction and modeling, evolution of metabolic pathways and regulatory systems, function and evolution of alternative splicing, functional annotation of genes and regulatory signals.