Available online at www.sciencedirect.com

SciVerse ScienceDirect

Genomics Proteomics Bioinformatics 10 (2012) 354–359

GENOMICS PROTEOMICS & BIOINFORMATICS

www.elsevier.com/locate/gpb

Method

TrFAST: A Tool to Predict Signaling Pathway-specific Transcription Factor Binding Sites

Umair Seemab [1,*], Qurrat ul Ain [1], Muhammad Sulaman Nawaz [1], Zafar Saeed [2], Sajid Rashid [1,*]

[1] *National Centre for Bioinformatics, Quaid-i-Azam University, Islamabad 44000, Pakistan*
[2] *Department of Computer Science, Quaid-i-Azam University, Islamabad 44000, Pakistan*

Received 4 April 2012; revised 10 June 2012; accepted 12 June 2012
Available online 2 December 2012

Abstract

Recent advances in the development of high-throughput tools have significantly revolutionized our understanding of molecular mechanisms underlying normal and dysfunctional biological processes. Here we present a novel computational tool, transcription factor search and analysis tool (TrFAST), which was developed for the *in silico* analysis of transcription factor binding sites (TFBSs) of signaling pathway-specific TFs. TrFAST facilitates searching as well as comparative analysis of regulatory motifs through an exact pattern matching algorithm followed by the graphical representation of matched binding sites in multiple sequences up to 50 kb in length. TrFAST is proficient in reducing the number of comparisons by the exact pattern matching strategy. In contrast to the pre-existing tools that find TFBS in a single sequence, TrFAST seeks out the desired pattern in multiple sequences simultaneously. It counts the GC content within the given multiple sequence data set and assembles the combinational details of consensus sequence(s) located at these regions, thereby generating a visual display based on the abundance of unique pattern. Comparative regulatory region analysis of multiple orthologous sequences simultaneously enhances the features of TrFAST and provides a significant insight into study of conservation of non-coding *cis*-regulatory elements. TrFAST is freely available at http://www.fi-pk.com/trfast.html.

Keywords: TrFAST; Transcription factor binding sites; *in silico* analysis; Signaling pathway; Pattern searching

Introduction

Non-coding DNA sequences have very important biological roles and exist in multiple types including non-coding functional DNA, *cis*-regulatory elements, introns, pseudo genes, repeat sequences, transposons and telomeres [1–4]. Among these, *cis*-regulatory elements are located at core as well as distal promoters and regulate gene transcription. These sequences might be present in 5′ or 3′ untranslated regions (UTRs) or within introns [5,6]. By definition, a transcription factor (TF, sequence-specific DNA-binding factor) is a protein that binds at a specific DNA site and regulates transcription [7]. The efficacy of TFs can be enhanced or reduced by binding of other regulatory proteins. Thus attachment of TFs to DNA affects the rate and efficiency of transcript initiation of a gene, either positively or negatively [8]. The specific sequence of DNA where TF binds and modulate the transcription of a gene is termed as transcription factor binding site (TFBS). Some TFs bind to DNA regions that are thousands of base pairs away from the gene they control and play significant roles in development, intercellular signaling, and cell cycle [9,10]. For this reason, multiple cancer phenotypes occur due to deregulation of TFs [11]. Among these are included the TFs belonging to the WNT pathway [12–17], Hedgehog pathway [18,19], NOTCH pathway [20], NF-κB pathway [21–23], MAPK pathway [24] and JAK/STAT pathway [25].

A variety of computational and experimental techniques have been employed to detect specific genomic regions

* Corresponding authors.
E-mail: u.seemab@gmail.com (Seemab U), sajid@qau.edu.pk (Rashid S).

http://dx.doi.org/10.1016/j.gpb.2012.06.007

bound by a TF [26,27]. However, the concurrent location of TFBSs from a set of sequences by a single click is lacking. By finding the TFBSs in multiple gene sequences, it would be possible to develop more robust descriptions of the TF binding properties. During this study, we developed a comprehensive tool TrFAST for detection of TFBSs which specifically bind to the regulatory regions of genes involved in a signaling pathway and to multiple sequences of greater length.

Results and discussion

Novel algorithms and protocols for pattern searching provide insight in the field of gene regulation. The demands on functionality and accuracy are challenging if without compromise in speed effectiveness. Graphical display of consensus TFBSs along with their locations in a whole range of input sequences makes TrFAST a unique tool. At the TrFAST initialization window menu bar, the user may open, browse or paste the input FASTA file containing the promoter sequences. The admin module provides the user access for addition or modification of signaling pathways and TFs. The bar, pie and pyramid charts help in analysing the respective promoter sequence in terms of the content of A, T, G and C. TFs selected by the user may be visualized in multiple species simultaneously. TrFAST can offer the analysis of TFBS in 20 sequences of greater length (maximum up to 50 kb). The efficiency and search-speed could be improved by reducing numbers of comparisons by algorithm.

TrFAST functionality is further enhanced by the explicit selection of signaling pathway, which helps to understand the functionality and role of TF according to a specific signaling pathway. Searching the TFBS according to the signaling pathway, integration of different genome browsers and access to diverse online tools or databases help to boost the tool's utility in analysis of different TFBSs. The prediction and accuracy rate of TrFAST is improved in a number of ways. It locates the exactly-matched positions of TFs or their combinatorial pattern at a given promoter sequence(s), while other TFBS searching tools produce a large number of false positives (FPs) at default threshold value, which reduce their efficiency and reliability [28]. It is tempting to speculate that, due to exact pattern matching, the chances of FP results are negligible in TrFAST. However, we cannot exclude the possibility of FP results. Until now, there is no algorithm available with full accuracy. The TFB search tools that use a weight matrix approach produce a large number of FP results [29]. A detailed overview of existing TFBS tools and TrFAST is compiled in Table S1.

Analyzing multiple orthologous promoter sequences in a single operation provides the user the benefit of identifying novel TFBSs, which have remained conserved throughout the defined time window. This conservation depicts the importance of the gene and its product for the survival of the organism. In order to test the success rate of TrFAST, we performed an upstream regulatory region analysis of the WNT family (**Figures 1** and **2**) by using the 7 kb intergenic sequences of *WNT-1*/*WNT-10B* and *WNT-6*/*WNT-10A* in four species including *Homo sapien*, *Mus musculus*, *Maccaca mullata* and *Danio rerio*. The presence of TFBSs for TCF-4, SOX9, n-myc, c-myc, TBP and TBF [30–32] (Tables S2 and S3) leads to the hypothesis that there may be conserved non-coding enhancer elements present in the upstream intergenic region of these WNT signaling.

Conclusion

TrFAST is a novel algorithm developed to facilitate searching for signaling pathway specific TFs in multiple data sets simultaneously. The features of TrFAST include sets of programming modules that increases its computational power as compared to all other tools. Hence, by highlighting the prediction accuracy issue using the modification in exact-pattern matching strategy followed by rescoring, the TrFAST tool improves upon many pattern searching tools with varying degrees of accuracy.

Methods

Data sources

In this study, we focused on the binding sites of TCF-4, LEF-1, HNF family, SOX family, TBF, TBP, ATF-CREB and ADF-1, P65, p50, c-REL, NF-kappa, E2F, c-jun, c-fos, NFY and AP-3, COUP-TF, CBF-1, CEBP, AP family, SP family and C-ETS, because of their major involvement in tumorigenesis. The potential TFBS for each TF are collected through extensive literature searches. As TrFAST involves TFs of specific signaling pathways, each pathway was carefully examined for TFs and their consensus binding sites. In order to incorporate the biological meaning of the study, promoter region sequences were downloaded from online genome databases including NCBI, Ensembl and UCSC, which have also been integrated in the TrFAST and assessed for TFBS. The efficiency of TrFAST was tested by using the promoter sequences of *Cyclin-D1, BRCA-1, BRCA-2*, *WNT-1, WNT-2, WNT-2B, WNT-3, WNT-3A, WNT-4, WNT-5A, WNT-5B, WNT-6, WNT-7A, WNT-8A, WNT-8B, WNT-9A, WNT-9B, WNT-10A, WNT-10B, WNT-11* and *WNT-16* genes [33].

TrFAST algorithm

The TrFAST algorithm has been designed as a modification of exact pattern matching algorithm (brute-force) [34] and applied to multiple sequences in a single step (**Figure 3**). It provides a user-friendly and graphical interface for visualization of TFs to perform comparative and detailed analyses (Figure S1). TrFAST includes the following steps:

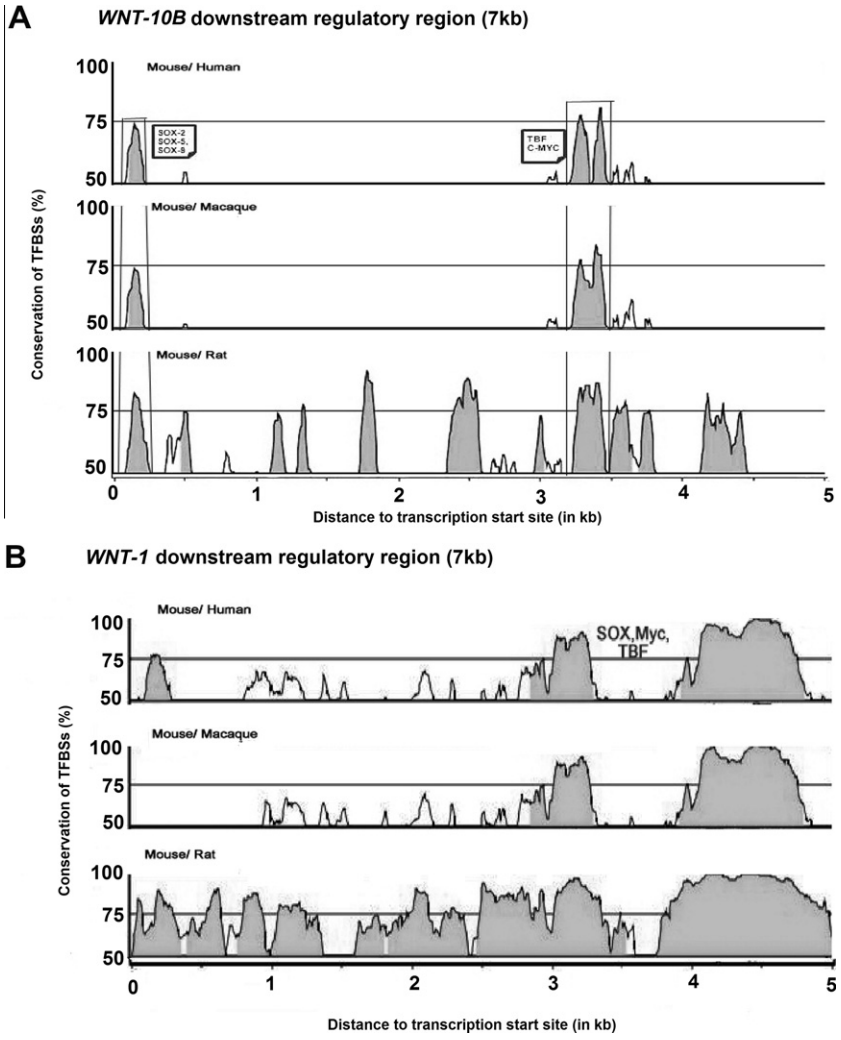

Figure 1 Conserved non-coding regulatory region analysis of *WNT-10B* and *WNT-1*

A. TFBSs for Sox family, myc family and TBF are conserved among *WNT-10B* in vertebrates. **B.** WNT signaling transcription factors were found conserved in downstream regulatory region of *WNT-1*. TrFAST was used to predict conserved TFBSs for regulatory elements in this region. MLAGAN was used to evaluate conservation of non-coding sequence (CNS) by comparing sequences among different organisms. Criteria of alignment were 100 bp window and 70% conservation cut off. Only the best results obtained within 5 kb sequence from transcription start site (TSS) (out of 7 kb) are shown in figure.

Step 1

The tool essentially takes FASTA sequence files as an input from the user and searches for possible TFBSs. The user can either upload the promoter sequence as a FASTA file or import the promoter sequences from web-based servers directly. In this step TrFAST ensures the proper FASTA file format and number of promoter sequences in the file.

Step 2

Upon uploading the promoter sequence(s), the desired signaling pathway is selected. The user can also customize the number of TFs or enter any new TF sequence as input.

Step 3

Multiple pattern searching is performed in a single effort by combinatorial searching. The possible number of permutations for any TF is calculated by

$$\prod_{j=1}^{3} n_j^{k_j} \tag{1}$$

Such that,

$$n_j^{kj} = n_1^{k_1} \times n_2^{k_2} \times n_3^{k_3} \tag{2}$$

where n_1, n_2 and n_3 represent different combinations of purine or pyrimidine in a given consensus sequence of TF; while k_1, k_2 and k_3 are the number of total n_1, n_2 and n_3, respectively.

$n_1 \in$ N, where N (any nucleotide) = {A, T, G, C}

$n_2 \in$ Z where Z = {H, B, V, D} and each item in H, B, V and D has exactly 3 base pairs, H $\in$ {A, T, C}, B $\in$ {T, G, C}, V $\in$ {A, G, C} and D $\in$ {A, T, G}

$n_3 \in$ X where X = {R, Y, M, K, S, W} and each item in X have exactly 2 base pairs, R $\in$ {A, G}, Y $\in$ {T, C}, M $\in$ {A, C}, K $\in$ {T, G}, S $\in$ {G, C} and W $\in$ {A, T}

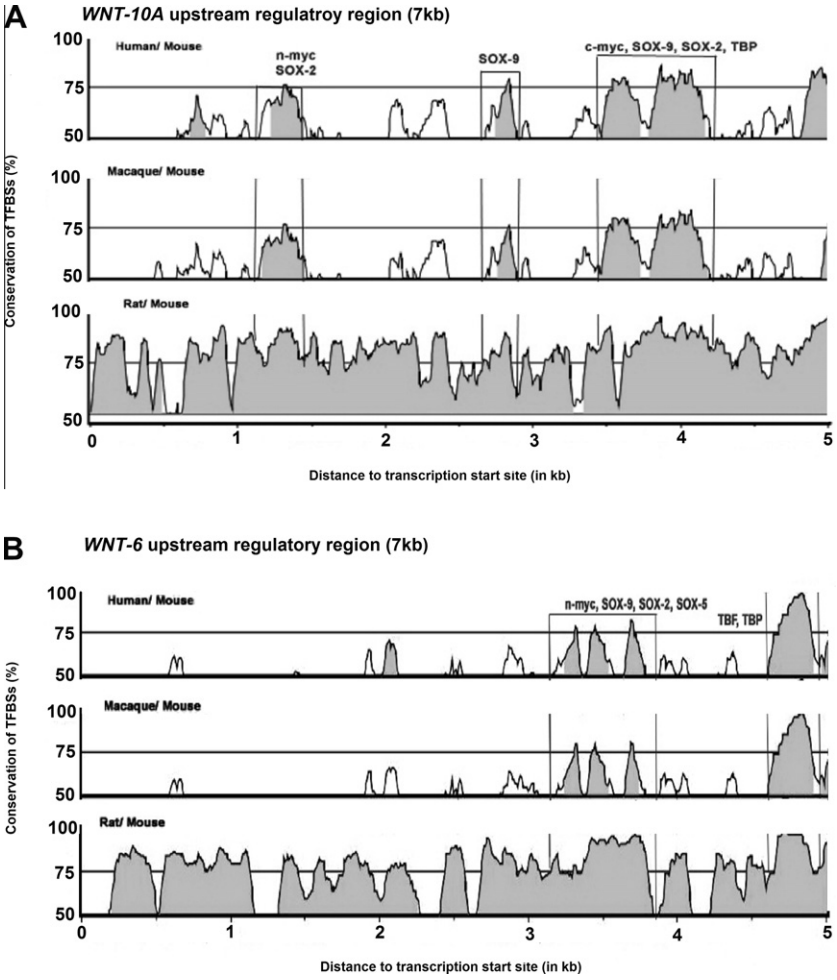

Figure 2 Conserved non-coding regulatory region analysis of *WNT-10A* and *WNT-6*
A. TFBSs for Sox family, myc family and TBF are conserved among *WNT-10A* in vertebrates. **B.** WNT signaling transcription factors were found conserved in downstream regulatory region of *WNT-6*. Analysis was performed similarly as shown in Figure 1.

Both Z and X make up the subsets of IUPAC ambiguous nucleotide codes, where **Z** denotes bases that code for three basic nucleotides with the remaining one excluded. For example, H codes for A, C and T but not G, similarly B, V and D do not code for A, T and C, respectively. In the same manner, **X** is a subset of bases which code for nucleotides grouped on the basis of purines, pyrimidines, presence of keto or amino group or weak and strong nucleotides. Permutation for multiple TFs in a single signaling pathway A can be computed as:

$$A = \sum_{i=1}^{n} \prod_{j=1}^{3} n_{ij}^{k_{ij}} \tag{3}$$

where n is the number of TFs in signaling pathway. As in TrFAST, the user can predict multiple TFs of a signaling pathway in a set of promoter sequences (up to 20 sequences) all together; the permutation of TFs calculated in all sequences will be given by,

$$A_{p_1} + A_{p_2} + \ldots + A_{p_{20}} = \sum_{j=1}^{20} A_{P_j} \tag{4}$$

In addition to extensive and manifold pattern searching, TrFAST is proficient in reducing the number of comparisons during exact pattern matching by implementing a novel strategy. In the case of N {**A, T, G, C**}, the computed index will move to Pattern_Index +1 without cross checking any nucleotide, thereby, avoiding repeated comparisons. Similarly, in the case of **H, B, V, D** (denoted by Z) in the pattern sequence, the algorithm will only check unacceptable nucleotides G, A, T and C, respectively, in the given promoter sequence. If these unacceptable nucleotides are found within the given sequences, the condition will be met and considered true. The TrFAST algorithm will consider this base pair's position as a mismatch and Pattern_Index will reset to the starting index resulting in Sequence_Index shift to +1. Thus this searching strategy reduces the number of comparisons to one third along with the computational time.

Step 4

In this step, the predicted TF binding positions (starting and ending indices) of an input sequence are stored for further use in a text file. Let ***l*** be the length of pattern to be

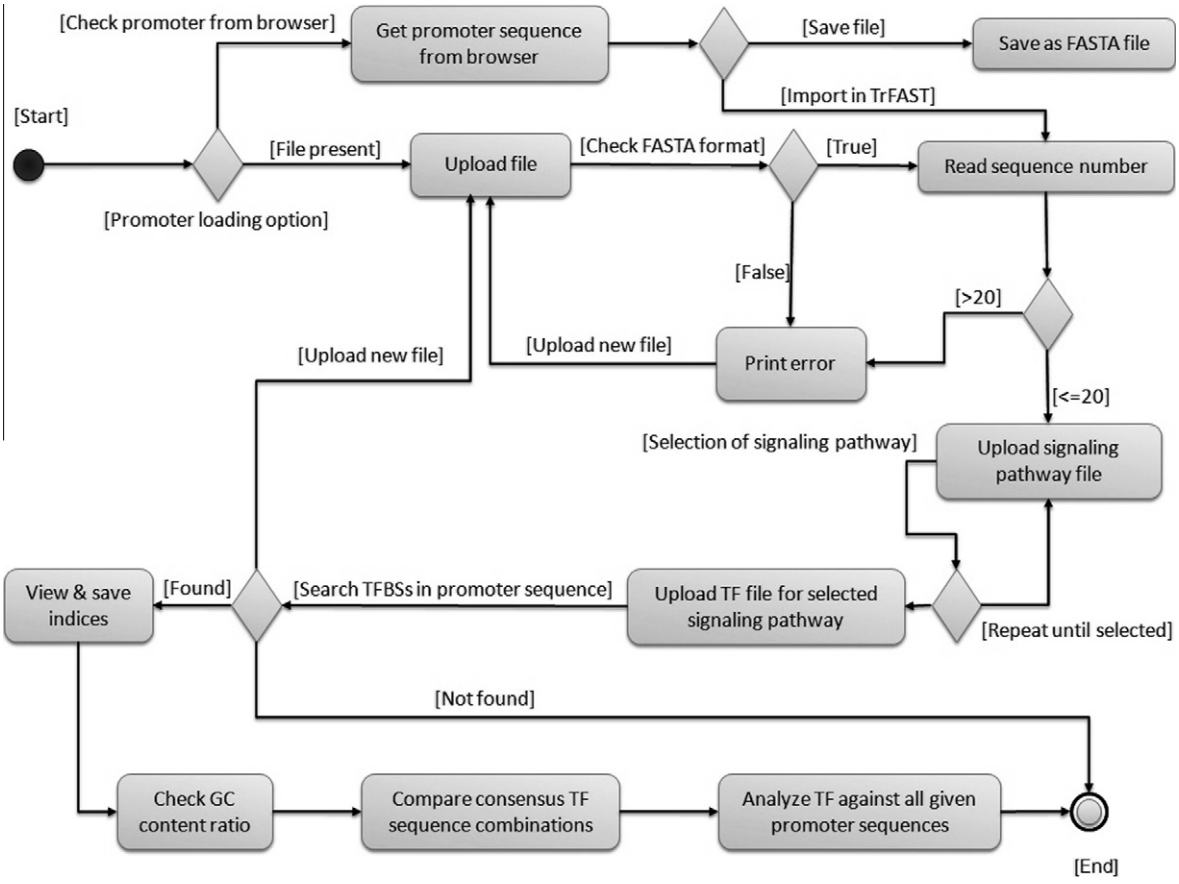

Figure 3 Work flow of TrFAST

The start event represents two-step sequence input module following the selection of signaling pathways and their corresponding TFs. Uploaded promoter sequences are analysed for the presence of putative *cis*-regulatory elements.

matched, which would always be >1. If the pattern exactly matches the input sequence, the starting index of the matched pattern, denoted by $\boldsymbol{fi}$ (where $\boldsymbol{i}$ is the current index of promoter sequence), and the ending index of the matched sequence denoted by $\boldsymbol{fi + l - 1}$ will be stored in a file along with their pattern sequences.

Step 5

During this step, the TrFAST algorithm inquires about existing matched patterns in Distinct_matched_Patterns array. If the pattern already exists, only newly found indices will be stored. Thus the new pattern will be added to Distinct_matched_Patterns array along with its indices.

Authors' contributions

US designed and implemented the algorithm. QA analysed the TFBSs at the promoter regions of proto-oncogenes. US drafted the manuscript with the help of QA and SN. ZS analysed algorithm. SR supervised the project.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Zahida Parveen and Nousheen Bibi for critical reading and editing of manuscript. This research is supported by Higher Education Commission, Pakistan (Grant No. 20-1493/R&D/09).

Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gpb.2012.06.007.

References

[1] Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo SW, et al. Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res 2007;17:839–51.
[2] Elgar G, Vavouri T. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends Genet 2008;24:344–52.
[3] Häsler J, Samuelsson T, Strub K. Useful 'junk': Alu RNAs in the human transcriptome. Cell Mol Life Sci 2007;64:1793–800.
[4] Carroll SB, Prud'homme B, Gompel N. Regulating evolution. Sci Am 2008;298:60–7.
[5] Rhoads RE, Dinkova TD, Korneeva NL. Mechanism and regulation of translation in *C. elegans*. WormBook 2006;Jan 28:1–18.

[6] Narsai R, Howell KA, Millar AH, O'Toole N, Small I, Whelan J. Genome-wide analysis of mRNA decay rates and their determinants in *Arabidopsis thaliana*. Plant Cell 2007;19:3418–36.
[7] Latchman DS. Eukaryotic transcription factors. London: Academic Press; 1998.
[8] Benz JR, Black HR, Graff A, Reed A, Fitzsimmons S, Shi Y. Valsartan and hydrochlorothiazide in patients with essential hypertension. A multiple dose, double-blind, placebo controlled trial comparing combination therapy with monotherapy. J Hum Hypertens 1998;12:861–6.
[9] Arnosti DN, Kulkarni MM. Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? J Cell Biochem 2005;94:890–8.
[10] Polak P, Domany E. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. BMC Genomics 2006;7:133.
[11] Andersen CL, Christensen LL, Thorsen K, Schepeler T, Sørensen FB, Verspaget HW, et al. Dysregulation of the transcription factors SOX4, CBFB and SMARCC1 correlates with outcome of colorectal cancer. Br J Cancer 2009;100:511–23.
[12] Libermann TA, Zerbini LF. Targeting transcription factors for cancer gene therapy. Curr Gene Ther 2006;6:17–33.
[13] Krauss G. Biochemistry of signal transduction and regulation. Weinheim, New York: Wiley-VCH; 2003.
[14] Van Es JH, Barker N, Clevers H. You Wnt some, you lose some: oncogenes in the Wnt signaling pathway. Curr Opin Genet Dev 2003;13:28–33.
[15] Amit S, Hatzubai A, Birman Y, Andersen JS, Ben-Shushan E, Mann M, et al. Axin-mediated CKI phosphorylation of beta-catenin at Ser 45: a molecular switch for the Wnt pathway. Genes Dev 2002;16:1066–76.
[16] Willert K, Brink M, Wodarz A, Varmus H, Nusse R. Casein kinase 2 associates with and phosphorylates dishevelled. EMBO J 1997;16:3089–96.
[17] Sakanaka C, Leong P, Xu L, Harrison SD, Williams LT. Casein kinase iepsilon in the Wnt pathway: regulation of beta-catenin function. Proc Natl Acad Sci U S A 1999;96:12548–52.
[18] Taipale J, Beachy PA. The Hedgehog and Wnt signalling pathways in cancer. Nature 2001;411:349–54.
[19] Kalderon D. Transducing the hedgehog signal. Cell 2000;103:371–4.
[20] D'Souza B. The many facets of Notch ligands. Oncogene 2008;27:5148–67.
[21] Gilmore TD. Introduction to NF-kappaB: players, pathways, perspectives. Oncogene 2006;25:6680–4.
[22] Brasier AR. The NF-kappaB regulatory network. Cardiovasc Toxicol 2006;6:111–30.
[23] Perkins ND. Integrating cell-signalling pathways with NF-kappaB and IKK function. Nat Rev Mol Cell Biol 2007;8:49–62.
[24] Seger R, Krebs EG. The MAPK signaling cascade. FASEB J 1995;9:726–35.
[25] Murray PJ. The JAK-STAT signaling pathway: input and output integration. J Immunol 2007;178:2623–9.
[26] Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, et al. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol 2012;13:R48.
[27] Whitfield TW, Wang J, Collins PJ, Partridge EC, Aldred SF, Trinklein ND, et al. Functional analysis of transcription factor binding sites in human promoters. Genome Biol 2012;13:R50.
[28] Won KJ, Ren B, Wang W. Genome-wide prediction of transcription factor binding sites using an integrated model. Genome Biol 2010;11:R7.
[29] Bulyk ML, McGuire AM, Masuda N, Church GM. A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in *Escherichia coli*. Genome Res 2004;14:201–8.
[30] Huang Z, Hurley PJ, Simons BW, Marchionni L, Berman DM, Ross AE, et al. Sox9 is required for prostate development and prostate cancer initiation. Oncotarget 2012;3:651–63.
[31] Packham G, Bello-Fernandez C, Cleveland JL. Position and orientation independent transactivation by c-Myc. Cell Mol Biol Res 1994;40:699–706.
[32] Amiel J, Rio M, de Pontual L, Redon R, Malan V, Boddaert N, et al. Mutations in TCF4, encoding a class I basic helix-loop-helix transcription factor, are responsible for Pitt-Hopkins syndrome, a severe epileptic encephalopathy associated with autonomic dysfunction. Am J Hum Genet 2007;80:988–93.
[33] Ain Q, Seemab U, Nawaz S, Rashid S. Integrative analyses of conserved WNT clusters and their co-operative behaviour in human breast cancer. Bioinformation 2011;7:339–46.
[34] Sheik SS, Aggarwal SK, Poddar A, Balakrishnan N, Sekar K. A FAST pattern matching algorithm. J Chem Inf Comput Sci 2004;44:1251–6.