Regulation of OCT-2

Farzon Nosiri

Sr. Software Engineer

Khujand, Tajikistan

Abstract*— *OCT-2, a member of the POU family of transcription factors, plays a critical role in regulating B-cell-specific genes and immunoglobulin transcription, essential for B-cell proliferation and differentiation. This study investigates the regulation of OCT-2 through promoter analysis and transcription factor identification. Using computational tools (PROMO and TESS) and DNA microarray data, five transcription factors—EGR2, Smad3, FOXJ2, TCF-7, and RXR alpha—were identified as potential regulators of OCT-2. Differential expression analysis suggests that RXR alpha, TCF-7, EGR2, and FOXJ2 may act as repressors in CBMI-Ral-STO cells, while Smad3 functions as an activator in RAEL cells, potentially mediated via the TGF-beta signaling pathway. OCT-2's involvement in Epstein-Barr Virus (EBV) latency regulation is also hypothesized, particularly its interplay with EBNA-1 in promoter regulation. These findings provide insights into OCT-2's transcriptional regulation and propose experimental pathways to validate its regulatory mechanisms. Understanding OCT-2 regulation may illuminate its broader implications in immune response and viral latency.

I. INTRODUCTION

Oct-2 is an octamer transcription factor which is part of the POU family, where the proteins have a well-conserved homeodomains. The POU proteins are eukaryotic transcription factors. They contain a bipartite DNA binding domain (POU domain) in which the POU homoeo and POU-specific regions form two subdomains that are both required for DNA binding but are held together by a flexible linker. The different members of the POU family possess a diverse range of functions, each associated with the growth of an organism. Oct 2 is necessary for octamer-dependent transcription of immunoglobulin and other important lymphoid-specific genes of B cells. The expression of Oct 2 is restricted to B cells 4and neuronal cells. It has an important role to modulate transactivation of IG promoter. In addition, Oct 2 participate in controlling B-cell specific genes which are used in proliferation and differentiation of CD20, CD79a and J. Chain (1). Oct 2 specifically binds to the octamer motif (5'-ATTTGCAT-3') (2) (3). The Isoform 5 of Oc2 activates the U2 small nuclear RNA (snRNA) promoter. (4) It also interacts with NR3C1, AR and PGR (5).

II. RESULTS

Before we start to the results, let's motivate our decision of taking two different software to predict TF binding sites. The reason behind this lies in collective intelligence, where a shared or group of intelligence, in our case predictions, that emerges from the collaboration of two or many individuals. (14) Also this is one of the ways doing collaborative filtering, when filtering information with the help of multiple agents and viewpoints, and data sources. (14)

Binding site prediction

In my work I have taken the promoter region of Lymphoid-restricted immunoglobulin octamer-binding protein (OCT2) which is encoded by POU2F2 gene to predict the Transcription Factors (TF) binding sites that can have influence on OCT2 regulation and expression. The size of the promoter region was selected from 2000 bp upstream and 200 bp downstream after the transcription start site which is in total make the sequence consisting of 2200bp long. Since the research is targeted on

human cell lines the transcription factors are selected among the human factors and binding sites also predicted for human organism. The predictions were performed with both software PROMO and TESS.

Input data

The promoter sequence with 2000bp for TESS, since TESS supports to process only the sequence of 2000 bp, the procedure was done the following way. First 2000 bp upstream from transcription start site was processed, then the other 2000bp with 1800bp upstream and 200bp downstream from transcription start site were analyzed. For Promo I have inserted the whole sequence of 2200bp, since there is no limit for input sequence. 966 transcription factors for TESS and 156 transcription factors for PROMO were selected to predict binding sites for them.

TABLE 1. The number of binding sites predicted for different similarity rates, from two software based only on TRANSFAC database.

Similarity rate	TESS	PROMO
1%	228	161
5%	249	331
10%	371	654
15%	1612	826
0%		

The results of binding sites (Table 1), according to different similarity rates are derived from only using TRANSFAC weight matrices and string matches for both software.

Unlike TESS PROMO resulted in more than 2 fold difference in each similarity rate change. The reason why TESS behaves differently is because TESS in addition to weight matrices provides string models. There are 223 binding sites for string models that bind with 100% similarity to the promoter region. Also one can say TESS has very few binding sites for weight matrices compared to PROMO. Explanation to this is that the cut off for TESS weight matrices are very strict. The minimum matrix similarity and core similarity are set to have equal values. For 1% dissimilarity rate minimum matrix core similarity and matrix similarity set to 0.99, for 5%, 10%, and 15% have been set 0.95, 0.90, and 0.85 respectively.

For the next prediction I used better and more justified input parameters. On the next table 2, the results are shown for all

99

available databases of TF binding sites. Both PROMO and TESS have not much overlaps of TF weight matrices and considering taking an overlap between them is not appropriate for our case. We are looking for any kind of TF from computational perspective taking into account the real biological behavior could possible bind to our promoter region. Even though let's have a look to the results of taking overlap between two software.
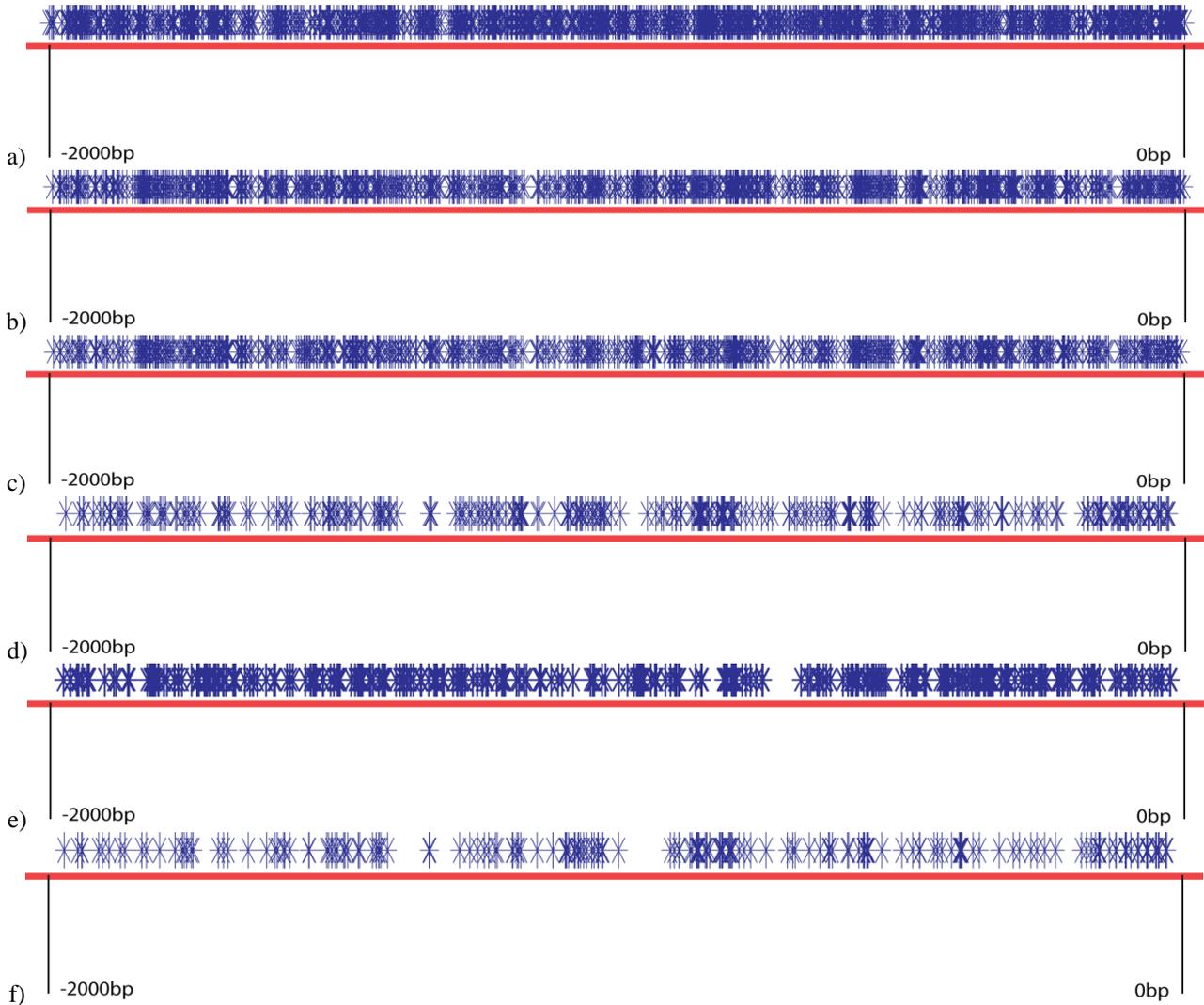

Figure 1. a) Binding site map of 15 % dissimilarity rate predicted by TESS, b) Binding site map of 15 % dissimilarity rate predicted by PROMO, c) Binding site map of 10 % dissimilarity rate predicted by PROMO, d) Binding site map of 10 % dissimilarity rate predicted by TESS, e) Binding site map of 5 % dissimilarity rate predicted by PROMO, f) Binding site map of 5 % dissimilarity rate predicted by TESS. Position 0bp is a transcription start site.
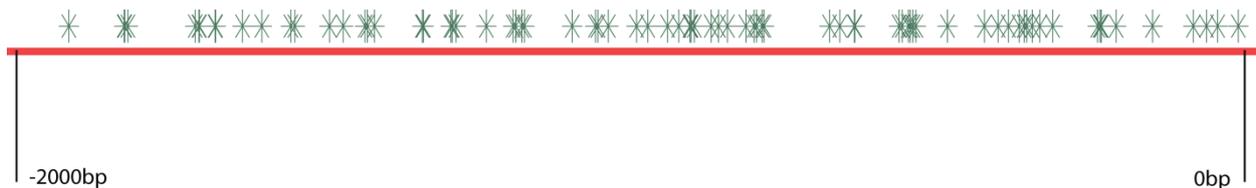

Figure 2. Overlap of binding sites, predicted with 10% dissimilarity rate, between TESS and PROMO.

To secure from different position numbering and also in a string length the overlap was taken with +/- 3 base pairs shift in position.

The predictions for the binding sites were conducted for different dissimilarity rates. The rates like 1%, 5%, 10%, and 15% were chosen. The numbers of predicted binding sites are shown on table 2.

1% input parameters:

For 1% dissimilarity rate the following parameters were given to TESS. There are 2 different kinds of parameters are considered. First initial settings for string matches are given the maximum allowed mismatch 1%, the minimum log-likelihood

100

ratio of 12 and minimum string length consisting of 6bp. The last 2 are default settings and recommended by TESS software.

TABLE 2. a) Table provides the name of the TF names the binding sites of which have been overlapped, b) comparative display of biding sites number that have been found in TESS (+200bp and -1800bp range) and PROMO (+200bp to -2000bp), c) shows the number of binding sites found with TESS for two ranges from -1800bp to +200bp and -2000bp to 0, also the last column shows the combined amount of binding sites from both ranges.

a)

Factor Name	Factor ID	Factor Name	Factor ID
Sp1	T00759	HOXD9	T01424
USF2	T00878	LEF-1	T02905
AP-2alphaA	T00035	MAZ	T00490
ATF-2	T00167	NF-1	T00539
C/EBPbeta	T00581	NF-AT1	T00550
c-Ets-1	T00112	NF-kappaB	T00590
c-Ets-2	T00113	NF-Y	T00150
c-Jun	T00133	p53	T00671
c-Myb	T00137	PEA3	T00685
c-Myc	T00140	Sp1	T00759
CREB	T00163	SRY	T00997
CTF	T00174	TCF-4	T02918
Elk-1	T00250	TCF-4E	T02878
GATA-1	T00306	TFIID	T00820
GATA-2	T00308	USF1	T00874
GCF	T00320	USF2	T00878
HOXD10	T01425	YY1	T00915

b)

	TESS 1800 200	PROMO
1	355	162
5	525	368
10	896	724
15	2277	914

c)

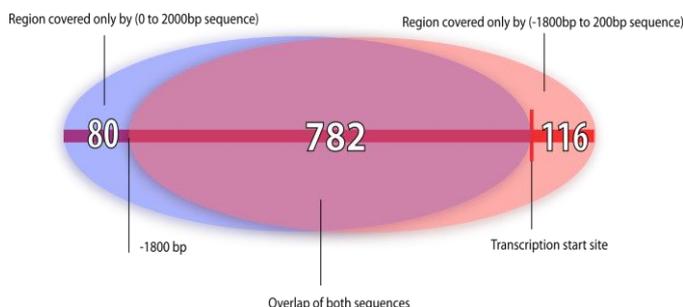|	TESS 1800 200	TESS 2000	Combined
1	355	331	381
5	525	492	565
10	896	862	976
15	2277	2109	2459

Figure 3. A graphical representation of two sets of predicted binding sites by TESS for regions from -2000 to 0 and -1800 to +200 which are combined together.

As for the weight matrices there are several factors are considered. I have chosen the scoring method according to matrix similarity and core similarity of weight matrix. The core similarity is given the value 0.9 which represents 90% of matches, and minimum matrix similarity 0.99 which denotes of 1% dissimilarity of whole sequence.

Other parameters like pseudocount were set by default as it is recommended by software. And also background probability of nucleotides in sequence was set by me and is taken from sequence distribution.
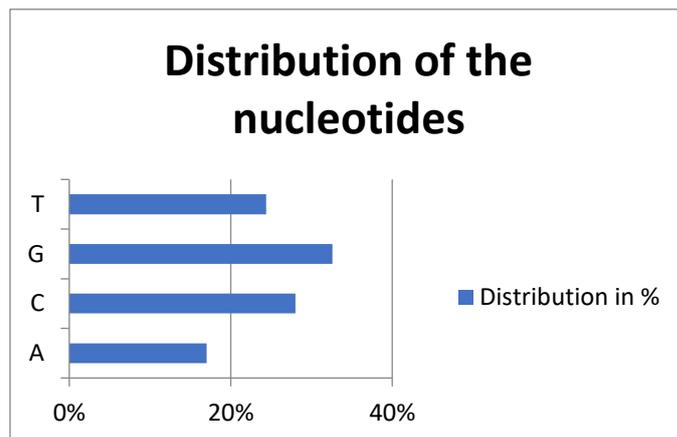

Figure 4. Distribution of the nucleotides over the given chain (sequence of 2200 bp).

On the table 3 the input parameters for different dissimilarity rate are shown. Parameters have been adjusted in order to optimize the output.

TABLE 3. Input parameters for different dissimilarity rate, both for string model and weight matrix.

Input parameters				
Maximum allowed mismatch	1%	5%	10%	15%
Minimum log-likelihood ratio	12	12	12	12
Minimum string length 6bp	6bp	6bp	6bp	6bp
Minimum core similarity (tc)	0.90	0.85	0.8	0.75
Minimum matrix similarity (tm)	0.99	0.95	0.9	0.85
Pseudocounts	0.1	0.1	0.1	0.1
Background probabilities	Seq. distr.	Seq. distr.	Seq. distr.	Seq. distr.

Matching With Microarray Data

Both software predicted and gave many binding sites. Next step was to compare and match the transcription factor list with affymetrix microarray data, which is an experimental data of transcriptome profiling of latency I and latency III cells. Some constraints were encountered with converting the TF ID to the gene id with the one in affymetrix data. The process was done manually one by one for each TF. As the result 141 TF from TESS and 81 from PROMO were successfully converted, which are in total make 222 TF. All data were converted to the Uniprot accession number and after to refseq nucleotide mRNA ID. Almost all accession numbers were successfully converted. Only one Uniprot accession number was not mapped. Totally they resulted in 356 unique Gene IDs which tells us that several of them have mapped to more than one ID.

Out of 356 TF from predicted binding sites 317 were matched with microarray data. Even if some IDs didn't match with microarray data there was an alternative ID from the same TF that have matched. As the result only 2 TF were not matched to the microarray. From top 20 matched TFs to the microarray data, which have differed significantly in expression between two latencies, 5 TFs that appeared to be biologically interesting and significant, and have good quality of binding site, have

101

been selected for further processing. In table 4 the reference list of 5 candidates are given.

TABLE 4. Five transcription factors listed on the table, which have been selected for further investigation. Here the expression according to affymetrix array and quality of binding site is shown.

Factor Name	CBMI-Ral-STO	RAEL	Quality
EGR2	1359.56	58.34	0.9
Smad3	232.78	1357.58	0.92
FOXJ2	1460.78	332.48	0.9
TCF	1750.95	26.13	1
RXR alpha	1058.1	103.14	1

The field "Quality" indicates the quality of binding site according to the method described in introduction. Since there are several binding sites along the promoter sequence predicted, here given the quality value of one the best matches.

Pathways

Next stage in the project was to find the pathways for selected transcription factors and look through the pathway up to the membrane in order to find and locate factors and proteins that could have influence the activation or inhibition of any selected 5 transcription factors. In order to elevate the pathway search for given factors, the current Affymetrix gene ids has been converted to KEGG gene ids. I have used DAVID id converter and Uniport's gene map to convert those ids that matches with KEGG database. One should make sure while converting genes they should belong to the same species, otherwise something wrong with input parameters for the converter. On table #X, one can find the reference table for each gene id in Affymetrix and ref.seq format converted to the KEGG database format.

TABLE 5. On table A the conversion of IDs from Affymetrix to Uniprot is shown. Table B shows mapped gene IDs to the KEGG database ID. This table provides IDs of different databases for selected 5 candidates.

A)

From	To	Species	David Gene Name
7984364	SMAD3_HUMAN	Homo sapiens	SMAD family member 3
7984364	Q9P0T0_HUMAN	Homo sapiens	SMAD family member 3
7984364	B7Z4Z5_HUMAN	Homo sapiens	SMAD family member 3
7984364	Q68DS8_HUMAN	Homo sapiens	SMAD family member 3
7984364	Q59F45_HUMAN	Homo sapiens	SMAD family member 3
7984364	B7Z9Q2_HUMAN	Homo sapiens	SMAD family member 3
8108050	TCF7_HUMAN	Homo sapiens	transcription factor 7 (T-cell specific, HMG-box)
8159127	RXRA_HUMAN	Homo sapiens	retinoid X receptor, alpha
8159127	Q6ZNL3_HUMAN	Homo sapiens	retinoid X receptor, alpha
8159127	Q6P3U7_HUMAN	Homo sapiens	retinoid X receptor, alpha
8159127	Q8NF63_HUMAN	Homo sapiens	retinoid X receptor, alpha
8159127	Q2NL52_HUMAN	Homo sapiens	retinoid X receptor, alpha
7933872	EGR2_HUMAN	Homo sapiens	early growth response 2
7953699	FOXJ2_HUMAN	Homo sapiens	forkhead box J2

B)

Name	NCBI	Uniprot	KEGG
EGR 2	NM_000399	EGR2_HUMAN	hsa1959
FOX J2	NM_018416	FOXJ2_HUMAN	hsa55810
RXRA	NM_002957	RXRA_HUMAN	hsa6256
SMAD 3	NM_005902	SMAD3_HUMAN	hsa4088
TCF 7	NM_003202	TCF7_HUMAN	hsa6932

III. DISCUSSION AND CONCLUSION

Our analysis revealed several potential transcription factors (TFs) that may play critical roles in regulating OCT-2 gene expression.

Firstly, we identified several TFs that are likely to act as repressors of OCT-2, including RXRα (6, 7), TCF7 (8), EGR2, and FOXJ2. These TFs exhibit higher expression levels in CBMI-Ral-STO cells compared to RAEL cells, suggesting an inverse correlation with OCT-2 expression.

Interestingly, we found evidence that the TGF-beta signaling pathway might activate OCT-2 expression. This is supported by the higher expression of SMAD3 (9), a key component of this pathway, in RAEL cells. However, it's important to note that EBV-encoded proteins can inhibit the TGF-beta signaling pathway (10, 11, 12), which could contribute to decreased OCT-2 expression in certain contexts.

Furthermore, our data suggests that the Wnt signaling pathway, particularly the canonical pathway involving TCF7 (8), may also repress OCT-2 expression. While core components of this pathway are expressed in both cell lines, potential differences in Frizzled receptors and Wnt ligands might influence pathway activity and its impact on OCT-2.

To experimentally validate these findings, we propose several approaches. These include manipulating signaling pathways by introducing TGF-beta ligands, activating the Wnt pathway, or treating cells with retinoic acid to observe the impact on OCT-2 expression levels. Additionally, techniques like ChIP-seq can be employed to directly determine the binding sites of these TFs on the OCT-2 promoter in vivo.

It is crucial to acknowledge the limitations of our microarray data analysis, such as the potential for unreliable results from low signal intensities (13). To address this, we suggest performing multiple microarray analyses to increase the reliability of our findings.

In conclusion, our results indicate that OCT-2 gene expression is likely regulated by a complex interplay of multiple TFs and signaling pathways, including the TGF-beta and Wnt pathways. Further research is necessary to fully elucidate these regulatory mechanisms and their implications in various biological processes.

REFERENCES

1. *Analysis of transcription factor OCT.1, OCT.2 and BOB.1 expression using tissue arrays in classical Hodgkin's lymphoma.* Garcia-Cosio M, Almudena Santon, Paloma Martin, Natalia Camarasa, Carlos Montalban, Juan F Garcia and Carmen Bellas. s.l. : Modern Pathology, 2004, Vol. 17.
2. *Transcription factor Oct-2A contains functionally redundant activating domains and works selectively from a promoter but not from a remote enhancer position in non-lymphoid (HeLa) cells.* Muller-Immergluck, M., Schaffner, W., Matthias, P. 5, s.l. : EMBO J, May 1990, Vol. 9, pp. 1625-34.
3. *A human protein specific for the immunoglobulin octamer DNA motif contains a functional homeobox domain.* Ko, H., Fast, P., McBride, W., Staudt, L. 1, s.l. : Cell, Oct 1988, Vol. 55, pp. 135-44.

4. *Promoter-selective activation domains in Oct-1 and Oct-2 direct differential activation of an snRNA and mRNA promoter.* Tanaka, M., Lai, J., Herr, W. 4, s.l. : Cell, Feb 1992, Vol. 68, pp. 755-67.
5. *Selective binding of steroid hormone receptors to octamer transcription factors determines transcriptional synergism at the mouse mammary tumor virus promoter.* Préfontaine, G., Walther, R., Giffin, W., Lemieux, M., Pope, L., Haché, R. 38, s.l. : J Biol Chem., Sep 1999, Vol. 274, pp. 26713-9.
6. *A novel octamer binding transcription factor is differentially expressed in mouse embryonic cells.* Koji Okamoto, Hitoshi Okazawa, Akihiko Okuda, Masaharu Sakai, Masami Muramatsu, Hiroshi Hamada. 1990, Cell, Vol. 60, pp. 461-472.
7. *The oct3 gene, a gene for an embryonic transcription factor, is controlled by a retinoic acid repressible enhancer.* H Okazawa, K Okamoto, F Ishino, T Ishino-Kaneko, S Takeda, Y Toyoda, M Muramatsu, and H Hamada. 10, 1991, The EMBO Journal, Vol. 10, pp. 2997-3005.
8. *TCF: transcriptional activator or repressor?* Bienz, Mariann. 10, 1998, Cell Biology, pp. 366-372.
9. *TGF- and the Smad signal transduction pathway.* Wrana, Arun Mehra and Jeffrey L. 2002, Biochem. Cell Biol., Vol. 80, pp. 605–622.
10. *Repression of Smad-dependent transforming growth factor-β signaling by Epstein-Barr virus latent membrane protein 1 through nuclear factor-κB.*

Mori N, Morishita M, Tsukazaki T, Yamamoto N. 5, 2003, International Journal of Cancer, Vol. 105, pp. 661–668.
11. *Down-regulation of the TGF-beta target gene, PTPRK, by the Epstein-Barr virus–encoded EBNA1 contributes to the growth and survival of Hodgkin lymphoma cells.* Joanne R. Flavell, Karl R. N. Baumforth, Victoria H. J. Wood, Gillian L. Davies, Wenbin Wei, Gary M. Reynolds, Susan Morgan, Andrew Boyce, Gemma L. Kelly, Lawrence S. Young, and Paul G. Murray. 1, 2008, BLOOD, Vol. 111, pp. 292-301.
12. *Epstein–Barr virus-encoded EBNA1 regulates cellular gene transcription and modulates the STAT1 and TGFbold italic beta signaling pathways.* Wood V H J, J D O'Neil, W Wei, S E Stewart, C W Dawson and L S Young. 26, 2007, Oncogene, pp. 4135–4147.
13. *Assessment of reliability of microarray data and estimation of signal thresholds using mixture modeling.* Musa H. Asyali, Mohamed M. Shoukri, Omer Demirkaya, and Khalid S. A. Khabar. 8, 2004, Nucleic Acids Research, Vol. 32, pp. 2323-2335.
14. Collaborative filtering. [Online] Wikipedia, the free encyclopedia, August 13, 2011. [Cited: November 12, 2011.] http://en.wikipedia.org/wiki/Collaborative_filtering.