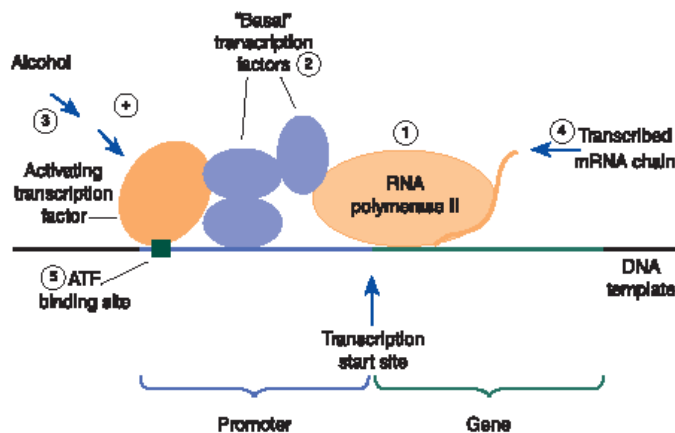


A Database of Transcriptional Regulation in the Human Genome



Student(s): Robert Marshall

Discipline: Bioinformatics

Supervisor(s): Dr. Robert Flegg, VBC, Dev Dehera, VPAC

Research Objective

The aim of this project was to create a database of transcriptional regulation in the human genome, with a web interface. It was implemented using Perl to retrieve and parse the data, MySQL to store it, and PHP to generate the content.

Motivation/Significance

Previously, transcription factors have only been studied on a single-gene basis. One would attempt to find out which transcription factors bound to a given promoter sequence. There are established methods for doing this, using either experimental or computational techniques.

However, with the completion of the Human Genome Project, it is possible to search every promoter region in the human genome for transcription factor binding sites, and collate the results. This would allow comparison of promoter regions across the whole genome, which is important as transcription regulates biological processes at a higher level than individual genes. To our knowledge, this has not been attempted before.

Many transcription factors are known to interact, but the processes are not well understood. Moreover, it is believed that there are many more undiscovered relationships between transcription factors. With a database of all transcription factor binding sites in the human genome, we can determine if certain binding sites commonly co-occur with a fixed distance between them. This is likely to indicate that, due to the folding of the DNA, transcription factors are close together in physical space, allowing them to interact.

It is also desirable to determine which genes are regulated by a given factor, or combination of factors. This could be useful when researching development, for example. If we know that certain transcription factors are only present during a given developmental stage, we can search for all genes which could be regulated by these factors, thereby identifying genes which may be used to regulate development.

Conclusion

We have created a database of transcriptional regulation in the human genome, as per our initial requirements. There have been some problems accessing and storing the data, but these can be easily rectified if the project is considered to be useful enough to purchase a subscription to TRANSFAC and some suitable storage.

The difficulties with the integrity of the data are more difficult to overcome. The high false positive rate is a product of the lack of data on experimentally verified transcription factor binding sites, and other ambiguities inherent in life, and is common to all current methods of computational transcription factor binding site prediction.

However, we believe that this project will be quite useful. The data we have used comes directly from TRANSFAC, the leading commercial database of transcription factors, so presumably it is of sufficient quality to be accepted. Based on discussions with Drs. Wolvetang and Young, we believe that the project will potentially be very useful to any biomedical researcher studying the regulation of development or other biological processes.