THE UNIVERSITY *of* EDINBURGH
School of Biological Sciences

Brongus: A Python-based Command-Line Program for Characterizing the Dynamic Transcriptional Functionality Programs of Promoters within the Genome of *Saccharomyces cerevisiae*

Student Exam Number: B080379

In partial fulfillment of the requirement for the Degree of Master of Science in Bioinformatics at the University of Edinburgh
2015/2016

Name of Dissertation Supervisor: Dr. Peter Swain

THE UNIVERSITY of EDINBURGH
School of Biological Sciences

Declaration of own work (Research dissertation)

All students must ensure the information below is included in their dissertation: simply copy and paste the text below. Each box must be ticked to show that the condition has been met, and it must be signed and dated - work will not be marked unless this information is provided.

FULL NAME: David Oliver Schlessinger

Matriculation Number: S1575558

Supervisor's Name: Dr. Peter Swain

Name of programme: MSc Bioinformatics

Submission Date: 23/08/16 No of Pages: 46 Word count: 12,686

I have read and understood the University of Edinburgh guidelines on plagiarism and declare that this written dissertation is all my own work, except where I indicate otherwise by proper use of quotes and references. I confirm I have:

- Clearly referenced/listed all sources as appropriate ☐
- Referenced and put in inverted commas all quoted text of more than three words (from books, web, etc) ☐
- Given the sources of all pictures, data etc. that are not my own ☐
- Not made any use of the essay(s) of any other student(s) either past or present ☐
- Not sought or used the help of any external professional agencies for the work ☐
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources) ☐

I understand that any false claim for this work will be penalised in accordance with the University regulations.

Signature: .....

Date: .....

Table of Contents

Abstract

Although there exists a number of web-based and standalone tools available to researchers interested in evaluating putative transcription factor binding sites in *Saccharomyces cerevisiae*, few of these tools provide the user with data pertaining to the total number of binding sites, their log-odd probabilistic scores, and their distribution per transcription factor-gene pair. Given the discovery of certain kinds of transcriptional functionality dependent on the amplitude and/or temporal profile of transcription factors, access to this data could prove to useful to researchers, particularly in characterizing the dynamics of specific expression profiles induced by a particular program of homo- or heterosynergy between transcription factors.

Brongus is a command-line program that could assist the elucidation of dynamic transcriptional regulatory systems that utilize the interaction of multiple transcription factors at particular concentrations by generating a tabulated, probabilistic data. This data's format distinguishes the potential binding sites in terms of their abundance, as well as collective and individual binding affinity scores per transcription factor-gene pair.

The binding affinity of a site is calculated using the sequence of nucleotides in a potential site and a transcription factor's motif, which is represented in the program as a Position Weight Matrix (PWM). For every transcription factor indicated by the user, the corresponding PWM is scanned along an assembly of promoter sequences to yield the number of potential binding sites (i.e. hits) between the transcription factor's motif and promoter's sequence, the log-odd probability score of each hit, the total sum of log-odd scores, and the average score per hit per transcription factor-gene pair. Conversely, the user could indicate one or more genes, so that their affiliated promoter regions are scanned by all PWMs in order to produce the same content and configuration of data as before. The user also has the option of uploading a text file with the names of multiple transcription factors or genes of interest instead of typing each one into the command line.

Histograms are produced for each analyzed transcription factor or promoter sequence so that the user can visually assess the distribution of the log-odd scores, thus allowing the user to infer where the transcription factor exhibits selective or promiscuous binding tendencies. And if the user indicates more than one transcription factor or promoter sequence, another table is produced with all the target genes or transcription factors that are shared by the multiple corresponding inputs.

Moreover, the possibility to set higher/lower thresholds of efficacious binding site affinity and different assemblies of genetic sequences to be uploaded allows for alternative potential binding sites to be evaluated within yeast. This information can then be interpreted or further processed by the user in order to infer whether the shared transcription factors with particular distributions of binding site affinity scores may be associated with a particular *cis*-regulatory module within a promoter. All this information can provide the user with a basis for further investigations in order to determine whether the predicted associations between the transcription factors or genes have any biological relevance.

Introduction

1.1 The Transcriptional Regulatory System of *Saccharomyces cerevisiae*

As a model organism, *Saccharomyces cerevisiae* (colloquially known as baker's yeast or budding yeast) has proved to be an indispensible tool for researching various biochemical mechanisms responsible for eukaryotic metabolism. As one of the first organisms to have its genome fully sequenced, a plethora of information has since become widely available in regards to its vital processes, including translation, transcription, proteomics, meiosis, DNA maintenance, and environmental interaction dynamics (Cherry et al., 2012). When external conditions change, cells must alter their metabolism to prevent the possibility of detrimental changes from occurring. Some examples of conditions that trigger specific reactions include excessive heat, which may cause the cell's proteins to denature, or nutrient dearth may cause the cell to run out of the raw materials or energy. In order to optimally respond to these shifts, signal cascades are relayed by distinct chemical messages that descend from plasma membrane receptors to the nucleus, usually via the phosphorylation of a series of intracellular secondary messengers. These chemical messages are subsequently interpreted by a group of proteins called transcription factors (TFs), whose function is to activate, promote, suppress, or halt the transcription of specific genes. This systematic alteration of the genome's expression profile in response to environmental changes can cause the cell to produce one set of proteins at a higher rate than another, carry out one set of metabolic activities over another, enter a different cell cycle phase, or induce sporulation. Such changes are pivotal for the survival of the cell (Brivanlou et al., 2002; Farkas et al., 2006).

In order to affect a gene's level of transcription, TFs must bind to particular sequences within the promoters regions of their gene target known as transcription factor binding sites (TFBS). Binding to these sites causes the three-dimensional structure of chromatin to become altered, which in turn causes the promoter region to become more or less accessible to RNA polymerase II (the enzyme for transcribing mRNA) and other transcription-promoting proteins. It is believed that a significant part of the eukaryotic

genome is either directly or indirectly affected by transcriptional regulation, including protein-coding and functional non-coding genes (Wray et al., 2003). There are an estimated 325 proteins associated with transcriptional regulatory activity in *S. cerevisiae,* of which 201 are designated as regulatory TFs (i.e. proteins that conditionally bind to specific motifs in the promoter regions of genes in order to affect their expression), while the others are identified as either transcriptional co-factors or as various transcription-associated proteins (Beskow et al., 2006). The number of highly characterized TFs in *S. cerevisiae* in one database, JASPAR, has been reported as 176 (Mathelier, 2015).

The activity of eukaryotic TFs can be classified as either constitutive or conditional. Constitutively active TFs alter the transcription of their target genes on a consistent basis, whereas conditionally active TFs only alter the transcription of their target genes in response to specific environmental conditions. The activation of conditionally active TFs is instigated either through the binding of certain ligands or by being dependent on the phosphorylation from the signal cascades from plasma membrane receptors. These signal cascade dependent TFs are then categorized either as resident nuclear factors or latent cytoplasmic factors. While resident nuclear factors are consistently localized to the nucleus, latent cytoplasmic factors translocate from the cytoplasm to the nucleus upon activation via phosphorylation (Brivanlou et al, 2002).

The systematic re-localization of proteins in response to certain environmental conditions has been detected to occur in many metabolic activities, including transcription (Tkach et al., 2012). The translocation of activated (i.e. phosphorylated) latent cytoplasmic TFs into the nucleus is facilitated by their binding to cognate nuclear envelope proteins, and their exit from the nucleus is typically mitigated by their deactivation, which is caused by dephosphorylation via phosphatases. However, this schema for the nuclear localization does not hold true for all latent cytoplasmic TFs, since the function and structure of different TFs likely mandate different mechanisms of activation, nuclear entry, deactivation, and nuclear exit. For example, when yeast cells are raised with access to an abundant supply of

inorganic phosphate, the latent cytoplasmic TF Pho4 is inactive, localized to the cytoplasm, and phosphorylated at four sites. If the supply of inorganic phosphate begins dipping to lower concentrations, then Pho4 is proportionally dephosphorylated and localized to the nucleus. When the concentration of inorganic phosphate decreases to a level that threatens the cells' survival, Pho4 then becomes fully dephosphorylated, localized to the nucleus, and active in transcriptional regulation (Lam et al., 2008). Additionally, a number of latent cytoplasmic TFs have been identified to constantly move in and out of the nucleus in a pulsatile fashion– this attribute will be discussed further below (Moll et al., 1991; Meyer et al., 2004; Dalal et al., 2014; Lin et al., 2015).

It is believed that once TFs entered the nucleus, they would search for target binding sites by facilitated diffusion, wherein the TF would move sporadically in three dimensions through out the nucleus until encountering some region of DNA and associating to it via non-specific interactions. The TF would then scan along the DNA via one-dimensional diffusion until the TF could bind to a site, else it would disassociate with the DNA and regress to three-dimensional diffusion (Dror et al., 2016).

This binding process to the TFBS is mitigated by the TF's DNA-binding domain, which is designed to have an affinity for distinct sequences of DNA. Such sequences are commonly referred to as motifs, and in eukaryotes, they can range from 5 to 30 nucleotides in length, with an average of about 10 nucleotides (Stewart et al., 2012). DNA-binding domains are classified based on their structure (e.g. leucine zipper, zinc finger, and helix-loop-helix folds), and can bind to DNA by determining the nucleotide sequence from the major or minor groove of DNA's sugar-phosphate backbone. However, some nucleotide sequences are "encoded" into the major or minor groove in the same way, leading to degeneracy in sequence identification (Dervan, 1986).

This degeneracy results in motifs with varying ranges of specificity, with some motifs allowing for more variation for the type of nucleotide present in certain positions versus others. Conventionally, researchers have represented this variation as a position frequency matrix (PFM), wherein the

nucleotides recorded in the sequences from confirmed binding sites are counted, or experimental data is converted to represent the counts of nucleotide per position (Stormo, 2013). These counts are then allocated into a matrix, where each position corresponds to the adenine, guanine, cytosine, or thymine base that occurred. The PFM can then be converted into a position proportion matrix (PPM), so that that the numbers at each position are proportional to the number of times a particular base occurs, and also all add up to one. The PPM can be further converted into a position weight matrix (PWM), also known as a position specific scoring matrix (PSSM), where the proportions are converted into their log-odd probabilities, as demonstrated in Figure 1. PWMs are integral in motif matching, where sequences of the genome (typically promoter sequences) are searched so that potential TFBSs can be located. The potential TFBSs are distinguished by multiplying the log-odd probability of the PWM's corresponding nucleotide relative to each position across the its length. So, if using the PWM from Figure 1, whichever sequence of 9 nucleotides in a given DNA region that yielded the highest number would be most likely to be TFBS (Stormo, 2013).

	1	2	3	4	5	6	7	8	9
A	3	6	1	0	0	6	7	3	2
C	2	2	2	1	0	2	1	1	2
G	1	1	6	9	0	1	1	4	1
T	4	1	1	0	10	1	1	2	5

	1	2	3	4	5	6	7	8	9
A	0.18	0.87	-0.91	-inf	-inf	0.87	1.02	0.18	-0.22
C	-0.22	-0.22	-0.22	-0.91	-inf	-0.22	-0.91	-0.91	-0.22
G	-0.91	-0.91	0.87	1.28	-inf	-0.91	-0.91	0.47	-0.91
T	0.47	-0.91	-0.91	-inf	1.38	-0.91	-0.91	-0.22	0.69

Figure 1. Example of Position Frequency Matrix (PFM) and its equivalent Position Weight Matrix (PWM) (made available by Guigo, 2003)

It was also assumed that the calculated log-odds probability with the highest number would have the highest binding affinity (Stormo, 2013).

Although motif matching might imply that the search for putative TFBSs in the promoter sequences is fairly straight forward, this could not be further from the truth. Many of the potential TFBSs detected by motif matching are actually not functional, mainly because either their location relative to the structure of the promoter sequences might preclude their functionality, or

because the compounding factors that operate synergistically to make the binding site become functional are not accounted for (Wasserman et al., 2004; Dror et al., 2016).

1.2 The Effect of Promoter Sequence Structure on Transcriptional Regulation

There are many structural features of *S. cerevisiae* promoter sequences that have been reported to dictate the positioning and activity of TFBSs– in particular, this includes the elements present in the core promoter, and the relative positioning of the nucleosome. The majority of TFBSs occur between 100 and 500 base pairs upstream of the open reading frame (ORF), which is also defined as the start site of translation (typically designated by the start codon of ATG). The maximum number of TFBSs for the average gene peaks at approximately 200 base pairs upstream (Guthrie and Fink, 2002, p. 249; Harbison, 2004). Beyond 500 base pairs upstream of the ORF, TFBSs occur much less frequently, although a few TFBSs have been located as far as 200,000 base pairs upstream (Wray et al., 2003). TFBSs are also generally rare in the region 100 base pairs upstream of the ORF, since this is where the core promoter is located. The core promoter is essential for the assembly of the pre-initiation complex (PIC), which recruits RNA polymerase II alongside other numerous cofactors in order to initiate transcription (Lubliner, 2013; Yang et al., 2007). Core promoters also include distinct elements that designate PIC assembly, such as TATA boxes, initiator elements (INR), other assorted elements (e.g. downstream promoter elements or CpG islands), or some combination thereof (Guthrie and Fink, 2002; Yang et al., 2007).

In general, upstream of the ORF is the location of the transcription start site (TSS); the length separating these two sites fluctuates greatly between different genes, and can even fluctuate depending on conditions for certain genes (Zhang et al., 2005). These inconstancies notwithstanding, the TSS consistently occurs adjacent downstream to INR-containing core promoters, while the assorted elements can occur either up or downstream of the TSS. However, in metazoan eukaryotes (e.g. arthropods, birds, or mammals), if a TATA box is present in the core promoter, then TSS starts 25 base pairs downstream. In *S. cerevisiae* and other related fungal species, the distance

between the TSS and the TATA-containing core promoter can occur anywhere from 40 to 120 base pairs upstream of the TSS (Yang et al., 2007; Tirosh, 2007).

This -100 base pair region can also be impacted by the presence of nucleosomes. Nucleosomes are structures within chromatin that package DNA, and they comprise of 147 base pairs encompassing an octomer of histone proteins. Some nucleosomes are classified as being “well-placed”, and are firmly positioned on DNA within the nucleus, while others are more “fuzzy”, and are reported to have dynamic localization (Lee et al., 2007). In general, promoter regions that are wrapped around a nucleosome are regarded to be non-conducive to transcriptional regulation as opposed to promoter regions depleted of nucleosomes, since nucleosome occupancy obstructs the access of TFs to their cognate TFBSs (Lam et al., 2008). This is also why euchromatin is typically considered transcriptionally more active than heterochromatin, since euchromatin has less nucleosome-occupied regions (Dror et al., 2016). And although a significant portion of the eukaryotic genome is involved with transcriptional regulatory activity, most of the genome is considered inaccessible to TFs due to chromatin positioning (Stormo, 2013).

However, genes whose promoters have a nucleosome present in proximity to their TSS (i.e. less than 100 base pairs upstream) tended to have a wider range of transcriptional plasticity (i.e. more dynamic response to stimuli), and were more likely to be associated with “fuzzy” nucleosomes with a high rate of histone turnover; their TFBSs were predominantly located in proximity to the TSS, greatly peaking at approximately 100 base pairs upstream, and occurring much less after 200 base pairs upstream. In contrast, genes whose promoters were occupied by nucleosomes at regions distal to the TSS (i.e. 150 to 400 base pairs upstream) tended to have less transcriptional plasticity and were more likely to be associated with “well-placed” nucleosomes; their TFBSs were distributed more conventionally, with a small peak around 200 base pairs upstream from the TSS, and more TFBSs located distally than proximally to the TSS (i.e. more than 200 base pairs upstream) (Tirosh, 2008). It is worth noting, that although less than 20% of genes in yeast are

reported to have a TATA box in the core promoter (Zhang et al., 2005), about half of the genes with nucleosomes present in the promoter region proximal to their TSS contain TATA box elements. Furthermore, the nucleosome positioning was retained in over 80% of genes in these two categories under heat shock conditions relative to normal conditions (Tirosh, 2008).

Interestingly, H2A.Z histone variants were consistently found with nucleosomes that localize distally relative to the TSS of promoters, as well as border nucleosome-depleted regions of DNA. This could be indicative that such histone variants are used in order to establish nucleosome free regions, although the mechanism underlying this process needs further investigation (Morse, 2007; Tirosh et al., 2008).

Furthermore, *S. cerevisiae* promoters that lacked a TATA box were found to be more “rigid” in structure than promoters containing a TATA box, thus making them less susceptible to nucleosome packaging, and more accessible to transcription factors. This is consistent with the previously described class of promoters whose nucleosomes are located distally from the TSS– they tended not to contain TATA boxes, and had a preponderance of TFBSs in the region less than 200 base pairs upstream of the TSS. Thus, regions of rigid DNA could dictate nucleosome positioning by being less conducive to nucleosome packaging than regions of flexible DNA, which indicates their heightened accessibility to TF binding. Additionally, regions of rigid DNA can manifest themselves into distinct structural features, such as, for example, an enhanced negative propeller twist (which is the rotation of an intra-base pair in relation to the Watson-Crick base pairing axis). Some TFs have been observed to bind preferentially to TFBSs located in (or near) one form of structural feature versus others. Given the aforementioned high number of TFBSs reported to be in or near regions of rigid DNA, it has been argued that DNA rigidity assists the facilitated diffusion of RNA polymerase II and TFs by providing a better “platform” to associate to, scan along, and then disassociate (Tirosh et al., 2007; Dror et al., 2016).

This can be further expanded by the observance of rigid DNA positioning at a consistent distance from the start codon in multiple species of

hemiascomycete yeasts, despite the marked divergence in sequence in between them. This could indicate that while nucleotide sequence may not be conserved, certain forms of DNA structure are conserved due to their functionality (Tirosh et al., 2007).

However, it is important to note that the structural feature a region of a DNA molecule conforms to is based primarily on its sequence of nucleotides. As a result, the nucleotides that border the motif of a TFBS have been observed to influence transcriptional regulatory activity. This influence is exemplified with the observations of some TFBSs with high-affinity motifs binding to their cognate TFs at rates comparable to TFBSs with low-affinity motifs– in other words, because the nucleotides flanking the TFBSs provided an adverse "platform" for the TFs to associate with, the transcriptional activity of the TFBSs with the high-affinity motifs was reduced (Dror et al., 2016).

Interestingly, the impact of bordering nucleotides can also cause TFs to differentiate TFBSs, even when their motifs are identical, so two TFs with the same DNA binding motif will selectively target TFBSs in different parts of the genome, since the environment of one TFBS is more favorable for one TF than the another. The physical limitation of distance excludes direct interactions between the TFs and neighboring sequences, which supports that their contribution to promoting TF binding is based on their influence on the shape. There is evidence to suggest that the optimal binding platform for TFs is determined (at least partially) by the percentage content of GC in the neighboring sequences (Dror et al., 2016).

In summary, if assessing a potential TFBS in a promoter region, one could also consider the following details: the location of the potential TFBS relative to the TSS, the composition of core promoter elements present in the proximal promoter region, the relative positioning of the nucleosomes, the structural feature of the local DNA molecule, and the nucleotides present in the sequences neighboring the potential TFBS.

And to draw further attention to the significance of promoter structure in influencing transcriptional functionality is the measured conservation of

nucleosome patterns and transcriptional plasticity between yeast and humans (Tirosh et al., 2008).

1.3 Synergistic Transcriptional Regulatory Activity

In eukaryotes, promoter sequences occur within intergenic regions of chromosomes, are generally reported as having lengths less than 1000 base pairs upstream of the TSS, although some promoter regions have recorded to be as low as 300 base pairs upstream from the TSS (Guthrie and Fink, 2002, p.115; Wray et al., 2003). The TFBSs located within promoters can also be further categorized as being either distal or proximal to the TSS, and research initially seemed to support the notion that proximal TFBSs were more likely to have functional significance versus distal TFBSs. But then it was revealed that the location of the TSS was not consistent within promoter sequences– although the TSS can be identified by a certain set of consensus sequences, there are genes with alternative TSS sites that are utilized in response to certain conditions (Zhang et al., 2005; Wassermann, 2004). This blurs the distinction between proximal and distal TFBSs.

Furthermore, the transcriptional activities of distal and proximal TFBSs are not always segregated from each other. Although there are examples of the conventional scenario in which one TF binds to one cognate TFBS in order to affect the single, concatenated gene's level of transcription, most transcriptional regulation requires multiple TFs (or cofactors) to concertedly bind to certain regions on the promoter in order to create a co-activator complex that instigates a particular form of regulation. These regions are referred to as *cis*-regulatory modules, and they are composed of multiple distal and/or proximal TFBSs because chromatin's three-dimensional structure can allow proximal and distal TFBSs to come into contact (Wassermann, 2004). Additionally, cis-regulatory networks are formed when the TFBSs from promoters affiliated with different genes are bound in a coordinated fashion in order to create specific functional clusters (Gao et al., 2013).

As mentioned earlier, eukaryotic TFs have an average motif length of about 10 base pairs; this is contrasted with prokaryotic average motif length of about 16 base pairs. While a shorter motif length does lead to a loss in binding specificity, it also makes the motif more robust, since longer DNA sequences have a higher chance of mutation, which can lead to a loss in binding affinity to the cognate TF (Stewart et al., 2012). The specificity of eukaryotic TFs is also bolstered by the previously discussed regions of DNA that are favorable to the binding of specific TFs through rigid structural features or preferable motif-flanking sequences. Another feature that augments such favorable regions is the presence of homotypic clusters, which comprise of multiple TFBSs adjacently grouped in the same genomic region that all bind to the same TF. Homotypic clusters can be considered one example of a *cis*-regulatory module, since their binding of multiple TFs can induce the concerted binding of multiple proteins (which can induce other processes such as chromatin remodeling), proportionally increase the gene's level of expression, or saturate the genomic region so that histones are displaced or occluded from binding. Additionally, the motifs of the TFBSs in homotypic clusters do not share the same level of binding affinity– rather, these adjacent motifs can act as an "affinity gradient", so once one TF associates with their genomic region, it scans along the motifs of higher and higher affinity until it can bind to the motif with the highest affinity. Supporting the notion that the composition of a genomic region can make the binding of specific TFs more conducive is the observation that homotypic clusters that have a high rate of similarity to the motif (Dror et al. 2016).

An overview of the pathways between various signal cascades to their target TFs expands on how the specificity of a transcriptional response clearly cannot rely solely on motif length, and how the relationship between the signal cascade, TF, and specific transcriptional response is not usually linear. Because signal cascades from many diverse sources can converge on the same TFs, the specificity of transcriptional regulation is mediated by transducing the information from signal cascades via two, sometimes overlapping mechanisms: hierarchical sub-networks of interacting TFs, and the amplitude and/or frequency dependent transcriptional regulatory programs (Hansen and O'Shea, 2015).

In the first mechanism, these hierarchical sub-networks are designed such that the TFs directly targeted by a signal cascade proceed to systematically regulate the transcription of genes that produce other TFs, which, in turn, can transcriptionally affect other TF-producing genes, and so on. Ultimately, this progression can elicit one or more particular outputs (i.e. transcriptional responses). These interactions have been characterized into 54 distinct sub-networks called origons, which operate concurrently to generate outputs tailored to the signal identity, as dictated by a combination of the pertinent input TFs, the intermediate TFs affected by the input TFs, and the lapse of time from the onset of the initial signal (Farkas et al., 2006). These sub-networks could also be described as *cis*-regulatory networks, since one TF systematically affects the transcription of a group of genes in order to produce a particular expression profile (which, in this case, is a particular sub-network of TFs).

But in the second mechanism, the desired transcriptional response is evoked as a consequence of two parameters: amplitude and temporal profile. The amplitude of a TF is defined by its concentration in the nucleus, while the temporal profile is determined by the frequency in which the TF enters and exists the nucleus. The temporal profile was previously referred to in regards to the nuclear translocation of latent cytoplasmic TFs, some of which were reported to exhibit a pulsatile behavior. The relationship between these TF's pulsatile translocation and transcriptional regulation is supported by evidence that measured changes in the frequency of the TF's pulses in proportion to shifts in environmental conditions (Dalal et al., 2014).

So, unless a certain concentration and pace is achieved, the transcriptional activity of certain genes will not be triggered. For example, Msn2, a latent cytoplasmic TF that responds to various stresses, induced different transcriptional responses from some of its target genes depending on whether it entered the nucleus in small or great quantities in conjunction with pulses that occurred at low frequency, high frequency, or were sustained over a duration of time. Furthermore, the functions of the target genes corresponded with their responses and the activation dynamics of Msn2

observed under disparate conditions (e.g. pulsatile nuclear localization in glucose starvation, but sustained nuclear localization under oxidative stress). As a result, promoters can be sorted into four non-discrete programs of activation dynamics: HF, HS, LF, or LS, where H stands for high amplitude, L stands for low amplitude, F stands for fast frequency, and S stands for slow frequency (Hansen and O'Shea, 2016).

In addition to the capacity for a TF to specify a certain transcriptional response with a certain temporal profile, it has also been shown that temporal profiles of multiple TFs can interact with one another in order to regulate a precise set of transcriptional activities. The presences of Msn2 and Mig1 in the nucleus were recorded to pulse sporadically under steady-state conditions, but when a decrease in the media's concentration of glucose was induced, their pulsing profiles became more distinct, with Msn2's pulsing pattern starkly not coinciding with Mig1's. This lack of synchronization led to the expression their shared regulatory target of GSY1, whose activation is mediated by Msn2, but whose repression is controlled by Mig1. The induction of other types of stress also caused the pulsing profiles of Msn2 and Mig1 to consistently react with or without synchronization relative to each other– oxidative stress resulted in Msn2 and Mig1's pulses to distinctly overlap, while heat stress resulted in non-overlapping pulses. Furthermore, increasing the intensity of the stress condition caused the frequency of the pulses to change proportionally. The mechanism behind this stress-dependent modulation of TF pulsing might be, at least indirectly, related to the catalytic factor of the phosphatase PP1, which can regulate the nuclear localization of Msn2 and Mig1 (Lin et al., 2015).

So, regulatory TFs can utilize a combination of any of the discussed mechanisms of transcriptional regulation so that the timing and/or level of expression from two or more target genes can be synchronized with the temporal profile and/or abundance of a certain TF. It is currently believed that the number of cognate TFBSs with high affinity determines the amplitude and timescale thresholds of promoters, as well as the local chromatin structure. However, the timescale thresholds of promoters are also ostensibly determined (at least partially) by their capacity for nucleosome remodeling

(i.e. their rate of histone modification, turnover, or displacement)– since more TFBSs could result in a higher duration of chromatin remodeling agents, which in turn increases the chance that certain, blocked TFBSs with different functionalities become exposed (Hansen and O'Shea, 2013; Todeschini et al., 2014; Goldschmidt et al., 2015).

It must be emphasized, that each TF does not share the same transcriptional regulatory behavior with every one of its target genes– rather, it is more accurate to posit that each relationship between a TF and the promoter of one of its target genes is characterized by a certain set of transcriptional regulation dynamics specific to the condition at hand. For example, as previously discussed, Msn2 responds to glucose starvation by translocating into and out of the nucleus in pulses with intensity-dependent frequency. But in response to osmotic stress, Msn2 translocates into the nucleus and stays there, with a duration proportional to the intensity of the osmotic stress (Hansen and O'Shea, 2013). Using information theory, researchers found that the different dynamic programs of TFs based on amplitude and frequency can be accurate in pinpointing the identity of the condition, but prone to errors in terms of specifying the intensity level of the condition. In order to limit the error rate in transmitting intensity-related information to the target genes, and thus evoke an apposite expression profile that is also proportional to the stress intensity, *cis*-regulatory modules can be designed to contribute the additional information (Hansen and O'Shea, 2015).

1.4 Correlations between the Affinity of Transcription Factor Binding Sites and Attributes of their Transcriptional Functionality

When numerous TFBSs are found on a promoter sequence (as observed in homotypic clusters), it can be indicative that their transcriptional functionality is contingent upon either homo- or heterosynergy (i.e. responding to the cooperative action of many copies of specific TF or many different specific TFs) (Todeschini et al., 2014). These clustered TFBSs have been reported to be associated with low-amplitude promoters, and as some promoters with clustered TFBSs were recorded to have high binding site affinities to their cognate TF, it could be surmised that the large number of high affinity TFBSs

greatly increase the chance of TFs successfully binding to their TFBS– this in turn means the low-amplitude promoter region does not require as many TFs for its transcriptional activity to be stimulated as a promoter with a high-amplitude threshold. Consistent with this idea are the reports of promoters with limited TFBSs for their cognate TFs that exhibit high-amplitude thresholds (Hansen and O'Shea, 2013).

But although high-affinity TFBSs may be present in a certain promoters, their functionality contingent may be on other factors that alter nucleosome positioning in order make them accessible to their cognate TFs. In a previously discussed scenario, the TF Pho4 is activated when inorganic phosphate concentrations become dangerously low. But even when inorganic phosphate concentrations are at moderate levels, Pho4 can still localize to the nucleus and bind to the promoters of its target genes. However, Pho4 selectively binds to the target gene PHO84 over PHO5 at a much higher rate, even though the promoters of these two genes have cognate TFBSs with low, medium, and high affinities. This is because the high-affinity TFBS in the PHO5 promoter is blocked by a nucleosome, while the high-affinity TFBS in the PHO84 promoter is located in a nucleosome-depleted region. But, there is a low-affinity Pho4 TFBS in the nucleosome-depleted region of PHO5's promoter that is functionally significant– when Pho4 binds to it, it can form a complex with ATP-dependent chromatin-remodeling proteins, causing the high-affinity TFBS to become accessible, and thus shifting the binding preference of other nuclear-localized Pho4 TFs from the promoter of PHO84 to PHO5. Despite the functional significance of this TFBS in PHO5's nucleosome-depleted promoter region, Pho4 does not readily bind to it as a result of its low affinity. Instead, this binding is instead mitigated by the presence of a co-activator Pho2, which has a cognate TFBS located in the same nucleosome-depleted region as the low-affinity TFBS. So for Pho4 to instigate the chromatin remodeling in the promoter region of PHO5, it must first form a heterodimer with Pho2, which, in turn, increases the chance of Pho4 binding to the cognate TFBS of low affinity.

However, Pho2 only interacts with dephosphorylated Pho4, and Pho4 is dephosphorylated only when the cells' concentration of inorganic phosphate

becomes dangerously low, which thus causes the expression of PHO5 to be specifically controlled by cell's supply of inorganic phosphate. However, it is worth noting that the formation of the Pho4-Pho2 heterodimer is not only control factor– as Pho4's nuclear concentration increases in response to decreasing levels of inorganic phosphate, its chances of binding to the low-affinity TFBS in the nucleosome-depleted region of PHO5 also increase (Lam et al., 2008). This scenario can be described as an example of a particular form of a *cis*-regulatory module– a heterotypic cluster that depends on the heterosynergy between TFs and their cognate TFBSs of varying affinities. And although it was previously mentioned that the timescale threshold of promoters has been linked to the capacity of the nucleosome to become structurally remodeled and increase their TFBSs' exposure, it has not yet been reported whether Pho4's transcriptional regulatory activity has pulsatile features associated with it in the same fashion as Msn2's.

The design and function of the homotypic and heterotypic clusters further serve to emphasize the significance high or low amplitude thresholds for promoters. The multiple TFBS copies proximally located to one another can be based in genomic regions that "funnel" the cognate TF to the apposite TFBSs, can displace histones upon saturation, or can induce the interaction specific co-factors in order to regulate the transcriptional functionality of a promoter region. Therefore, it could be extrapolated that one purpose of these *cis*-regulatory modules is to decrease (or, in some cases, increase) the amplitude threshold of a promoter region. Additionally, the number of potential binding sites for a single TF is positively correlated with the level of gene expression (Dror et al., 2016).

And despite the correlation between stronger binding site affinity and higher level of transcriptional activity (Sharon et al., 2012), many TFBSs with relatively low affinity have been found to bind with TFs– although some of these interactions have been found although to be functional, a great many more require further study (Lam et al., 2008; Tanay, 2006). In one study that utilized some TFs from metazoan eukaryotes, some TFBSs with calculated low-affinities were bound just as often as TFBSs calculated to have high affinities (Badis et al., 2009). As previously illustrated, the binding of some

TFBSs with low affinity can be mitigated by dimerization between two TFs (e.g. Pho4 and Pho2), whose cooperation can increase the binding affinity by increasing their collective motif length. A longer motif length strengthens specificity, and also allows for more potential complementary interaction between the DNA binding domains and the DNA. It should also be noted that some eukaryotic TFs increase the affinity of their TFBSs by having multiple DNA binding domains present (Todeschini et al., 2014).

It has been previously shown that heterotypic clusters can have many potential TFBSs that border each other. But the function of such clusters can be supplemented when the TFBSs overlap each other, which can cause certain sites to have multiple functional variants (i.e. it can be bound to by different TFs, and thus be induced to have different transcriptional functionality). This is because one binding site can bind to only one TF at a time, and, if the motifs of multiple TFs with similar targets overlap one another, then usually the TF with the highest affinity and/or the highest concentration will bind. This causes the site to have different transcriptional functionality depending on the conditions (Wray et al., 2003). Eliciting the apposite transcriptional function from these multi-functionality sites can be further specified by the selective guiding of TFs by favorable genomic regions– the set of TFs that are conducive to the region with the heterotypic cluster might change depending on the conditions (Todeschini et al., 2014; Dror et al., 2016).

It can therefore be surmised that some correlations could be drawn between the number and calculated affinity scores of TFBSs on a given promoter, the binding attributes of the potential TFBSs (i.e. whether their transcriptional functionality is contingent upon homo- or heterosynergy), and the structures of the promoter and localized chromatin in order to determine the adjoining gene's properties of transcriptional functionality.

1.4 Overview of Selected Databases and Resources for TFBS Assessment

In order to assist research dealing with the complexities and massive amount of data related to transcription in yeast (as well as other metabolic processes), scientists have compiled a number of databases and resources

that draw from an extensive set of research and data. The most prominent of these resources is the Saccharomyces Genome Database (SGD), which is an immense database containing a virtual encyclopedia describing the functions, locations, and details for every gene, protein, and lipid in *S. cerevisiae*, as well as the interactions between them and their supporting literature. Since its inception, the data and tools SGD have constantly been updated and maintained, and each feature of the *S. cerevisiae* genome (from the reference strain S288C) is assigned a unique SGD identification number (SGDID). These feature types include the protein-coding ORFs, but also include feature types that occur within the intergenic regions (i.e. non-ORF regions), such as autonomic repeating sequences (ARS), tRNA coding genes, rRNA genes, Ty elements, telomeric elements, and centromeric elements (Guthrie and Fink, 2002, p.471). These chromosomal features that have all been verified in literature– features whose existence require further investigation are designated as dubious or uncharacterized. This unique identification system is useful for distinguishing between genes that may share common or alternate names with one another, and also serves as basis by which the gene, its protein product(s), and the associated functions can be interlinked. In addition to the SGDID, ORFs and tRNA genes are further assigned a feature name that is consistent with the systematic nomenclature conventions. The database can be accessed through most Internet browsers, and offers a profusion of different services and information, including (but not limited to) BLAST, gene ontology, customizable genome browsers and downloads, sources to pertinent, up-to-date literature, a community forum, lists of experimental methods and reagents, and many additional resources (e.g. pertinent links to external, specialized databases) (Cherry et al., 2012).

Through the YeastMine search tool, the summary pages for TFs in the SGD database also offer a wide range of transcription-related data, and include relevant information from two external databases: JASPAR and YeTFaSCo. The TF's corresponding entry in the JASPAR database is linked through a JASPAR accession number. This entry contains additional information related to the transcriptional activity of the TF, such as the position weight matrix of its motif, the class of its DNA binding domain, its regulatory

interactions with other genes and/or proteins, its source in literature (via a Pubmed identification number), the locations of its documented binding sites, and more. The confidence level of the binding site sequence motif(s) associated with the TF are calculated through YetFaSCo using the affinity scores, and are given along with links to the TF's corresponding entry in YeTFaSCo (Costanzo et al., 2014).

There are a number of different databases available today dedicated to illustrating the transcriptional activity in the yeast genome, such as JASPAR, YeTFaSCo, Yeast Promoter Atlas (YPA), and SwissRegulon (Pachkov et al., 2007; de Boer and Hughs, 2011; Chang et al., 2010; Mathelier et al., 2015). Each database utilizes different sets or combinations of core data and algorithms for assessing potential TFBSs, is tailored to address disparate aspects of transcription, and yields outputs in various formats which may or may not be accessible to further processing. Not only do most of these databases have links to pertinent, additional information back on SGD and/or protein databases, and they also interlink between each other.

The standalone usage of such databases reveals their discrepancies in their motif searching algorithms and data content. For example, most of aforementioned databases are capable of scanning at least one DNA sequence for potential TFBSs by a set of non-redundant, curated PWMs of TF motifs. In JASPAR, the sequence location, name of potential TF, and corresponding motif sequence for each putative sites is given, which are all ranked based on their score relative to the calculated affinities of the PWM (Mathelier et al., 2015).

In YeTFaSCo, the sequence locations and name of potential TFs are given, but the putative sites are ranked based on their position along the sequence, and their scores calculated from the percentage match between the motif's PWM and potential binding site. This difference in potential TFBS scoring alludes to how divergent the intrinsic structures of the databases are– JASPAR has one entry per TF and its associated motifs, while in YeTFaSCo, each motif has its own individual entry, which means that a single TF can be represented multiple times within the YeTFaSCo database (although the

entries can be consolidated if they have the same TF name). The motifs are distinguished from each other based on their PWM, the type of experiment from which they were determined, and their confidence score, which is used in order to quantify the motif's functionality in the genome. The confidence score for each motif was calculated using an amalgamation of information, including experimental data (such as measured probe intensity), concurrence with other studies, and gene ontology enrichment. Unlike JASPAR, results can be downloaded in various text formats (de Boer and Hughs, 2011).

In Yeastract, the name of the potential TF, its consensus motif sequence, the location, and a literature reference are given for each putative TFBS, but not the calculated affinities or even motif match percentages, since the matches are determined based on whether a particular sequence matches with a TF's PWM or not– the binding affinity of a potential TFBS is not quantified. However, one can upload multiple sequences at once and compare their hits, a feature which is not readily available in other similar databases (Teixera et al., 2014).

In contrast to the previous databases, the Yeast Promoter Atlas (YPA) database does not offer the example service of scanning a given DNA sequence for potential TFBSs. Instead, YPA is dedicated to amassing the structural data of promoters that may influence transcription. The promoter regions for 6603 genes are stored along with their lengths, location within the yeast genome, the positions of their TSS and ORF's start and stop site, and the quantity, sequences, and locations of TATA boxes present. It also details the number and locations of TFBSs, the relevant references in literature confirming them, additional information of the cognate TF, the length and sequence of TFBSs, the number of references that confirm a detected change in the gene's expression when the cognate TF of a TFBS is mutated, the proportion of positions in each TFBS occupied by nucleosomes under different conditions (e.g. in vitro, in rich media, in galactose, or in ethanol), the proportion of positions classified as rigid DNA in each TFBS, a visual sequence map, and links to external databases. Although not all information is available for every promoter region, this database offers a user-friendly

interface capable of querying multiple promoters or TFs, and also allows the output to be downloaded in text format (Chang et al., 2010).

In short, different databases incorporate different types of information about transcriptional regulation and offer diverse analytical tools. Thus, the user must merely determine which information they are interested in, choose which outputs suits their conceptual and practical needs best, and then pick the corresponding database(s).

1.5 The Objective of Brongus

The Brongus code was developed in order to easily produce sets of malleable, comprehensible data that characterized potential TFBSs based on their number of occurrences and respective calculated affinity score per TF-gene pair, as well as the scores' total sum and average. The current databases do not make this kind of information readily available, which thus can impede the characterization of genes whose promoters may be characterized by a certain amplitude threshold or affiliated with a particular *cis*-regulatory module based on interaction between multiple TFs and TFBS with varying levels of affinity.

This code is designed to encourage further analysis of the output, and it is highly recommended that the output be compared and/or used in conjunction with pertinent information from other databases, particularly with YPA.

The ability to accurately identify the locations the TFBSs targeted by a particular TF within genomes has greatly advanced since the advent of high-throughput technologies combined with other experimental techniques (e.g. ChIP-seq). Furthermore, this information is accessible on a variety of databases. However, such experiments can only study one TF at a time, which can make inferring the TF's participation in heterosynergistic *cis*-regulatory modules and networks difficult. Furthermore, as previously discussed, the locations of bound TFBSs are contingent upon the conditions of the cell, which can further obfuscate the transcriptional regulatory system the TF(s) adhere to. Given the large extent of cooperativity between TFs and

other cofactors required for specifying transcriptional regulatory responses, and given the advantage in accurately predicting TFBSs in genetic sequences *in silico*, the development of tools to characterize the properties of such *cis*-regulatory modules would be of highly significant value (Wasserman et al., 2004; Stormo, 2013; Gao et al., 2013).

Although more development will be required before it can be implemented on a wide-scale basis, Brongus could be used to aid researchers hoping to elucidate the transcriptional functionality of homotypic and/or heterotypic clusters in the promoters of genes by examining their calculated binding affinities to one or more particular TFs.

Methodology

2.1 The Development, Basic Set Up of the Code, and SGDConvertermodule

It was decided that the program be written in Python because of its practicality, widespread support, and because of the access to the various modules of Biopython (Cock et al., 2009). The program was designed to run as a platform-independent, command line program through a terminal or Python shell. The command line interface was also set up using the Argparse module (Davis, 2015). It was designed to include one required argument (out of four mutually exclusive options), one conditionally required argument, and two optional arguments. The contents and purposes of these arguments are discussed below.

Chief among the code's potential implementations was its capacity to aid in the detection of synergistic transcriptional functionality between certain TFs and the promoters of their target genes, particularly in *cis*-regulatory modules with homo- or heterotypic clusters, promoters with "affinity gradients", or promoters with TFBSs with varying levels of binding affinity and different transcriptional functionalities. Thus, researchers would not be limited to looking at the binding affinity scores of the TFBSs for a single TF or on a single gene promoter. Instead, researchers could indicate one or more TFs or gene promoters in the command line, and the applicable analyses would be carried out.

At the most basic unit, the code would calculate the binding affinity score between a given nucleotide sequence (i.e. the promoter) and the PWM of a TF by calculating their log-odds probability of binding at each position in the sequence. Thus, the program would need to utilize at least two sets of data in order to conduct the analyses: an array of high-quality PWMs of TFs and an assembly of genetic sequences (i.e. the promoter regions of genes). Since the PFMs of yeast TFs from the JASPAR database (stored as .pfm files) have been used in numerous studies, are readily utilized by a number of Biopython modules, are highly expanded upon in the SGD, and are the most recently updated, they were chosen to be the array of TF PWMs. Although Biopython also included modules that could build PWM motifs from other formats (including MEME or TRANSFAC), the JASPAR format also stores a lot of meta-information per each PFM, and can easily be converted into a PWM form (note: referred to as PSSM in the Biopython code), which is required for calculating the binding affinity scores based on log-odds probability (Mathelier et al., 2015).

However, not all 176 PWMs available for *S. cerevisiae* from JASPAR were used in Brongus' code. The PWM with the JASPAR accession number MA0330.1 was excluded from the array of PWMs because it was composed of two separate proteins of MBP1 and Swi6 (Pachkov et al., 2007). Because the code is currently designed for calculating the binding affinity scores for one gene/TF pair, its inclusion in the code kept yielding errors during analyses. A future development of Brongus will thus have to accommodate this discrepancy.

Ultimately, the assembly of genetic sequences to be utilized was chosen to be the promoter sequences of the yeast genes from Yeastract. Each of the promoter sequences in Yeastract was given as exactly 1000 base pairs in length, which could accommodate the TFBSs of most genes, as discussed above. Its .fasta format also made its parsing easy and practical in the code, since Biopython includes a function for converting .fasta files into dicts. However, the possibility of utilizing other assemblies of genetic sequences is available, as discussed below.

It should also be noted that the promoter regions of genes will be referred to merely as genes from this point on, unless otherwise noted. This is mainly for the sake of abbreviation but also for clarity– the promoter regions of genes will be indicated in the command line and code by their affiliated gene's name.

Additionally, the indication of the TFs or genes in the command line also required a precise identification system within the code. Given that some genes and TFs have no common name or share common/alternate names with one another, it was clear that the code would have to be designed to avoid possible confusion. Further complicating the matter was the naming of the TFs' .pfm files, which were based on JASPAR's database accession numbers. It was ultimately decided that the SGDIDs of the gene or TF would be used as the primary naming basis since they were easy to sort, and moreover, were used to index the SGD_features.tab from the Saccharomyces Genome Database website. The table from this file not only included corresponding standardized feature names, but it also included common and alternate names (when available), genomic location, the feature type of the gene (e.g. ORF, tRNA gene, etc.), verification status, and a brief description of the gene.

Therefore, a sub-module SGDConvertermodule to have an IDConverter object that was designed to convert the user-indicated name to one specified SGDID pertinent to the gene or TF, regardless of whether the user indicated a common name, feature name, or even an SGDID. Furthermore, SGDConvertermodule would create a Python dict called SGD_dict built using SGDIDs as keys and objects as values, wherein each object contained pertinent details from the SGD_features.tab, including common name, feature name, alternate names, feature type, verification status, and a brief description of the gene. For the sake of narrowing the scope of the analyses, only two feature types of genes were included in the SGD_dict: ORF genes (protein-coding genes) and tRNA genes. If the user indicated the name of a TF or gene that could not be matched with a SGDID, an error would result. This would serve to limit the analysis not only to genes from *S. cerevisiae*,

but only genes with feature types registered as ORFs or tRNA genes within the SGD database. Besides allowing the user flexibility in providing the names of TFs or genes of interest, this system would keep the names and attributes of the TFs or genes consistent within the code itself, allowing the different modules to give each other information without the errors of misidentifying genes that could share common/alternate names. This would allow the PWM of a TF and its gene properties (e.g. common name, feature name, brief description) to be quickly corresponded with one another. And finally, the attributes of the SGDID object can be called at ease, making it very practical to pull out the relevant information whenever needed (e.g. the corresponding feature name or brief description).

In addition to the .pfm files of the yeast TFs, two more files were downloaded from the JASPAR database: TFMATRIX.txt, and Annotation.txt, both of which were modified to remove all data not relevant to the information about *S. cerevisiae* (and subsequently had their names changed to SaccCereTFMATRIX.txt, and SaccCereAnnotation.txt, respectively). These files contained information that allowed conversion of the TFs' JASPAR accession numbers to their common names (note: these common names were then converted into the corresponding SGDIDs) as well the TF's corresponding Medline number. These Medline numbers could then be used to access the source in literature from where the .pfm was calculated by adding the number to the end of the Pubmed URL in the format: http://www.pubmed.com/medlinenumber.

The scripts, data, and outputs of program were designed in order to be easily packaged in a .zip folder and run through a terminal: a general folder containing all the scripts of the code and two sub-folders for Data and Output. The Data sub-folder would contain a folder of the TFs' .pfm files from JASPAR as well as the .fasta file of Yeastract promoter sequences, SGD_features.tab, SaccCereTFMATRIX.txt, and SaccCereAnnotation.txt.

The scripts included in Brongus are the SGDConvertermodule, change_keys_to_SGD, motifscorecompiler, motifscorereader, and

common_assembler. A general overview of the input, data processing, and output of Brongus can be seen in Figure 2.

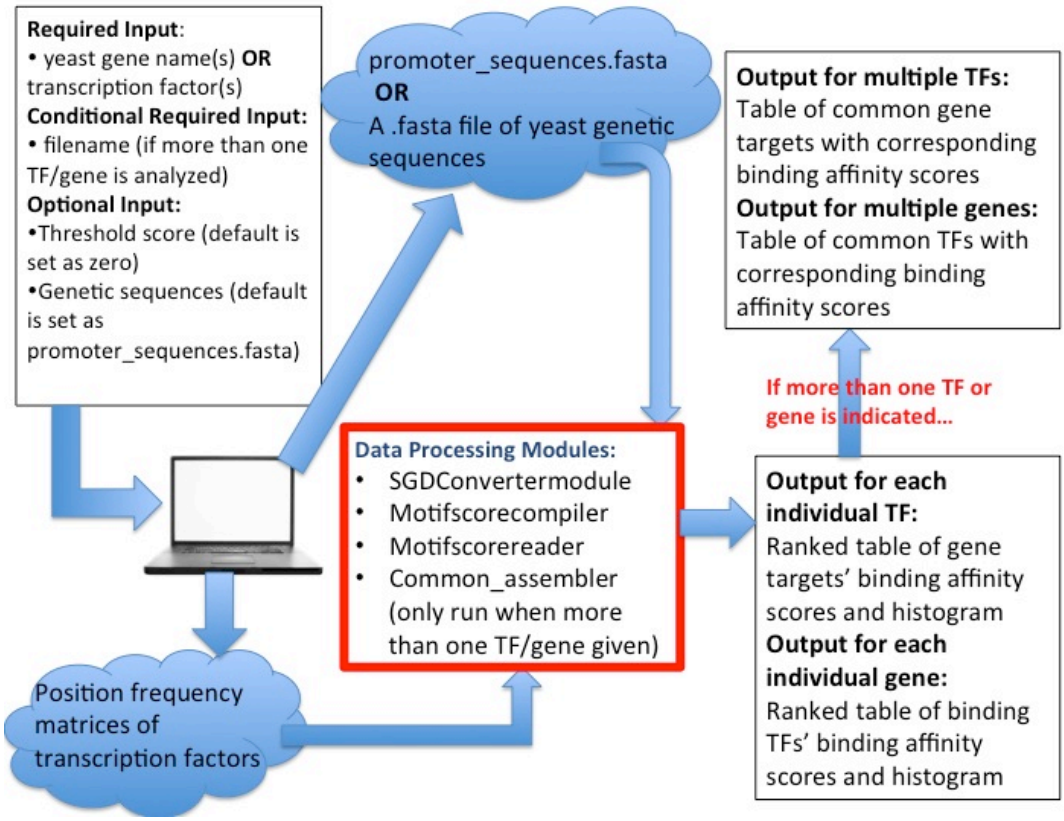

Figure 2. General Overview of Workflow in the Brongus Code.

The two clouds represent two main sources of information that are utilized in order to build the output (i.e. the JASPAR .pfm files that scan the Yeastract promoter sequences). If the user indicates one or more TFs, then their PWMs will scan against all promoter sequences with motifscorecompiler. But if the user indicates one or more genes, then their affiliated promoter sequences will be scanned by all PWMs, also with motifscorecompiler. Potential hits are set as scores greater than zero, and the hits of these analyses are then stored as .csv tables and histograms by motifscorereader. When multiple TFs/genes are indicated, an additional .csv table is generated with all the shared gene targets or TFs by common_assembler.

2.2 The motifscorecompiler sub-module

In the first module of motifscorecompiler, the array of PWMs and assembly of genes are both processed to become Python dicts with the corresponding SGDIDs as the keys of the dict. However, the values of the dict built from the .fasta file (referred to in the code as promoter_sequences) are Biopython sequence objects, while the values of the dict built from the array of PWMs (referred to in the code as the motif_dict) are Biopython motif objects, which are converted from their original PFM format into a PWM format (note: this format is referred to as PSSM in Biopython). Thus, the genetic sequences that will be scanned by the PWMs of the TFs are organized in their own separate dictionaries, but identified by their corresponding SGDIDs. Furthermore, it should be noted that not all 6781 sequences from the original Yeastract promoter_sequences.fasta file were used– only genes whose names could be matched back to a SGDID item located in the previously described SGD_dict. Because the promoter_sequences.fasta file named the sequences either with their common or feature name (and also variously with or without underscores), and included genes with disparate feature types including ORF genes, tRNA genes, and others, formatting had to be done in order to convert the sequences' names to SGDIDs. This formatting was carried out by the change_keys_SGDID script. Ultimately, only 6666 sequences of the original .fasta file were used, all other promoter sequences whose names that could not be converted were thus excluded from the assembly of sequences (i.e. the promoter_sequences dict) and subsequent analyses.

After the dicts for the PWMs and promoter sequences are set up, the scores for each position on a given sequence are calculated using the PWM of a given TF. If the user indicated one or more TFs, then the designated TFs would scan against all sequences in the promoter_sequences dict in their forward and reverse complement forms using the Biopython motif module. Conversely, if the user indicated one or more genes, then the designated sequences of the genes would be scanned by all PWMs from the motif_dict in their forward and reverse complement forms. In both cases, a list is

generated of all probabilistic log-odds score generated by the interaction between the PWM and each position along a sequence.

If any of the scores calculated fell below an indicated threshold, they were ignored, while the scores above the threshold (i.e. the “hits”) in both the forward and reverse directions were accumulated together and stored as a list in a new dict’s value with a corresponding key. Depending on whether the user’s input was a TF or gene, the key of this dict would either be the gene’s or TF’s SGDID, respectively. If any particular gene/TF pair failed to yield any scores above the threshold (i.e. no “hits” were detected) at any positions above the threshold, then this pair is excluded from the dict of hits between each gene/TF pair. This generated dict was subsequently serialized as a Python pickle, and then stored for later use in building the output. All pickle files were stored in the “Pickles” folder of the Output folder, and were named with the format of either “tf_SGDID_scores.pickle” or “gene_SGDID_scores.pickle”, where SGDID was the unique identifier, and tf/gene indicated what type of analyses were run.

2.3 The motifscorereader sub-module

Once all pickles are built and stored for each TF or gene indicated by the user, they can then be individually opened up and read in by the second module, motifscorereader. After being read in, the list of “hits” are processed to calculate their total sum and their average. A data table in the .csv format is then generated in order to represent the total recorded interactions between a single TF and the gene sequences or a single gene sequence and the TFs. The indices of the table is based on SGDIDs of the gene or TF, while the columns are set (from left to right) for the corresponding feature name, common name, number of hits, total sum of binding affinity scores, average of binding affinity scores, the list of binding affinity scores (sorted from lowest to highest), and a brief description of the gene’s or TF’s function. The data tables generated for the genes are distinguished from the TFs since they include a column containing the Pubmed URL of the TFs. The .csv tables are ranked from highest to lowest total sum of binding affinity scores,

and then accordingly exported either to the “TF_CSV” or “Gene_CSV” folder in the Output.

Additionally, a new list is generated that accumulates every score calculated from every recorded “hit”. This list is then used to create a histogram that shows the distribution of log-odds scores, which can be used to see where the peaks of the binding affinity scores occur. These peaks can then be used to determine whether a certain TF has selective or promiscuous binding patterns, or whether a certain gene has many potential binding sites with high or low affinity. The histograms are then accordingly exported to the “TF_Histograms” or “Gene_Histograms” folder in the Output. Examples of these histograms can be seen in Results section.

It should be noted that, although the SGDID names are used throughout the code to keep track of the pertinent TFs or genes, the names provided by the user are also kept and used in the filenames the .csv tables and are also included in the filenames of the histogram. This allows the user to keep track of genes or TFs with names they are more familiar with.

2.4 The common_assembler sub-module

As previously stated, this code’s main purpose was to detect the potential interactivity between TFs and genes using the correlations in the calculated binding affinity scores. This is where the third module, common_assembler, comes into play. The pickles of each indicated TF or gene are successively read in, opened up, and processed to determine their corresponding number of hits, total sum of binding affinity scores, and average binding affinity score. Simultaneously, a list is generated per pickle file that accumulates the SGDIDs that were recorded to have at least one hit with the TF or gene at hand. These lists are then run through a logical set operation that determines the common elements shared between two or more sets. This yields a list of common genes or common TFs shared by the corresponding input TFs or genes.

Finally, a table is set up in which the indices are based on the SGDIDs of the common elements. The number of columns in the table expands in proportion to the number of inputs, so that each gene or TF has its own corresponding column for number of hits and total sum of scores. Columns for the feature name, common name, description, and Pubmed URL (in the case of an input with multiple genes) are also included, and when completed, the .csv table is accordingly exported into the “TF_CSV” or “Gene_CSV” folder in the Output.

The columns for average binding affinity score and list of scores were not included in order to prevent the .csv tables from becoming too cluttered and having decreased legibility. But since the module motifscorereader ran for each individual input, the pertinent information is available on the gene’s or TF’s individual .csv file. The filename of the .csv table must be provided by the user as an argument before the analyses start in order for Brongus to function.

2.5 The Set Up for The Arguments in the brongus_user_interface

As previously mentioned, the Argparse user interface is designed to run Brongus with one required argument (out of four mutually exclusive options), one conditionally required argument, and two optional arguments. The group of four mutually exclusive options is comprised of one TF type argument, one gene type argument, one TF filename type argument, and one gene filename type argument. The TF type and gene type argument both require the names of at least one TF or gene of interest, respectively, to be typed into the command line, each separated by one space. The TF filename and gene filename arguments also require the names of at least one TF or gene of interest to given– however, the names are given in simple .txt files uploaded to the working directory with each TF or gene separated by a new line. This was designed so that a researcher interested in analyzing a large group of TFs or gene promoters, but not interested typing in each of their names individually into the command line, could simply upload a text file with the desired names. However, it is crucial that the names of TFs and genes are

placed in separate text files– otherwise the output data would be faulty or a possible error would occur.

As mentioned before, arguments for TF and gene will not yield an error unless they are given at the same time and/or the indicated gene names cannot be converted into their corresponding SGDIDs by the IDConverter tool. Additionally, the indicated name for the TF is under more constraint, since it must match a corresponding SGDID as well as a JASPAR .pfm file.

The arguments for TFs and genes were made mutually exclusive because the names for the genes of TFs and the TFs themselves (i.e. as proteins) had the potential to be interchangeably referred to, which could lead to some errors in the analysis. Thus, the code was designed to make the user explicitly choose whether they wanted to analyze the interactions between one or more TFs and the genes or vice versa.

The conditionally required argument was briefly touched above– if a user provides more than one TF or gene, then providing a filename for the .csv table that assembles the common gene targets or TFs becomes a required argument. Initially, the code was developed without this argument, so if the user gave more than one TF or gene, the user-inputted names were concatenated and given as the filename for the .csv tables with shared genes/TFs. However, this filename would become arduous to read once the user provides more than 5, 20, or even 100 inputs. It also might be difficult to differentiate the filename from other .csv tables with similarly long names. So, it was decided for the user to provide a filename that they could be familiar with as a conditionally required argument.

2.6 The Significance of The Threshold and Genetic Sequences Argument

Although the final two arguments are set to be optional, as they both have defaults, they represent two highly significant features of the Brongus code, as their modulation could potentially provide the user with cogent information for characterizing previously unidentified synergistic transcriptional functionality between multiple TFs and genetic sequences.

The process of calculating the log-odds probability score of a PWM binding to a given sequence has already been illustrated above, and the conventional model posited that an overall lower calculated score was correlated with a lower binding affinity between the sequence of nucleotides at the site and TF, which in turn, indicated a lack of transcriptional functionality at the particular site. However, it has since been shown that the conventional model can not always account for the numerous functionally relevant TFBSs with low binding affinities, such as those described in the "affinity gradients" or the chromatin-remodeling sites. Additionally, the potential TFBSs located by two different but similar PWMs can greatly diverge from each other depending on how low the threshold value is set (Vorontsov et al., 2013). Thus, researchers wanting to see how the TF/gene pair's distribution of log-odd scores and number of hits expands or contracts in response to threshold changes could make use of threshold adjustment. For example, if the promoter sequence of a particular gene was suspected to have a high amplitude to transcriptional regulation by one or more TFs, then lowering the threshold would allow for the observance of a few isolated, low to moderate affinity potential TFBSs, which could aid in the characterization of the promoter sequences' transcriptional regulatory program. Furthermore, the .csv table with the shared TFs or gene targets can be used in order to see if one or more TF/gene interactions shares a similar set of high or low binding affinity scores.

The second optional argument provides the user the capacity to analyze an alternate assembly of genetic sequences by the PWMs instead of Yeastract's promoter_sequences.fasta. This design gives the user the possibility of narrowing down their analyses of binding affinity scores with one or more TFS to only a few promoter sequences, rather than the default 6666 promoter sequences. Another advantage would be that the histograms generated from said analyses would only focus on distribution of log-odd scores in a much smaller scope, which could promote the detection of certain binding trends.

Additionally, as the argument's name "genetic sequences" implies, the user has the possibility to upload other genetic sequences besides promoter regions. Because although there is a propensity for TFBSs to be located in the promoter region, some TFBSs have been located outside of the promoter region, such as in the affiliated gene's ORF and even introns. Although their functional purpose is disputed, it is possible that they represent alternate systems of transcriptional regulation (Li et al., 2001; Borneman et al., 2007; Jun et al., 2010). Furthermore, the various mutants of genetic sequences could be potentially uploaded in order to assess the changes in putative TFBS locations– a topic which has been investigated many times, and which is even featured on the YPA database (Wray et al., 2003; Chang et al., 2011). This is one of the reasons why the promoter regions have been referred to as genes in this code and in the section of this paper– in order to accommodate for the possibility that genetic sequences other than promoter regions might be analyzed.

There are only three constraints for utilizing a different assembly of genetic sequences: they must be in a .fasta format (since the code converts the contents of a .fasta file into a Python dict), the file must be placed in the Data folder of Brongus, and their IDs must be recognized by the SGIDConvertermodule– else, they are excluded from analysis by the code.

2.7 Comparison of Results with Other Databases

The information from the .csv files can be used in conjunction with or merely compared to other databases, particularly Yeastract and YPA. For instance, a user can retrieve a list of the feature names from their .csv tables, enter them into the "Search Regulatory Associations" page on Yeastract, and see whether their indicated TF(s) or gene(s) interactions were also detected. The user can also look at additional information on YeastMine under the Regulation tab. Additionally, genes whose distribution of log-odd scores show potential to have "affinity gradients" or some other *cis*-regulatory module may have their promoter structure checked on YPA to look at the promoter's nucleosome positioning, rigid DNA presence, and number of TATA boxes (if present). This information, in turn, could be used to support

possible characterizations of the promoter's transcriptional regulatory program.

Results

3.1 Individual and Shared TF and Gene .csv Tables

The .csv files generated for each individual TF or gene can opened and further processed in commonly accessible data-processing programs, such as Excel or RStudio. (Figure 3, Figure 4). The brief description of the genes or TFs can serve as a cursory guide for excluding biologically irrelevant hits or correlating certain functional groups.

	Gene Feature Name	Common Name	Number of Hits	Total Sum of Scores	Avg. Score per Hit	List of Scores	Gene Description
S000005360	YNR077C		71	447.310862	6.30015299	[0.093017071, (	Protein of unknown func
S000003988	YLL065W		51	262.777016	5.15249051	[0.2242616, 0.2	Dubious open reading fr
S000001831	YFL063W		50	260.90606	5.2181212	[0.093017071, (	Dubious open reading fr
S000002951	YDR543C		47	236.078614	5.02294924	[0.2242616, 0.2	Dubious open reading fr
S000028709	YOL166W-A		44	235.952092	5.36254756	[0.093017071, (	Protein of unknown func
S000004945	YMR326C		43	228.054526	5.30359362	[0.11573715, 0.	Dubious open reading fr
S000002173	YDL015C	TSC13	42	207.355839	4.93704378	[0.2242616, 0.5	Enoyl reductase; catalyz
S000028536	YCR108C		37	201.419198	5.44376211	[0.2242616, 0.2	Putative protein of unkn
S000001830	YFL064C		31	196.686259	6.34471802	[0.2242616, 0.5	Putative protein of unkn
S000004330	YLR338W	OPI9	57	182.552756	3.20267993	[0.074280508, (	Dubious open reading fr

Figure 3. Example output of the TF Met32.

By selecting on a cell in the "List of Scores" column, users can examine the distribution of log-odd scores on every promoter region that was calculated to have at least one hit with the TF Met32.

	TF Feature Name	TF Common Name	Number of Hits	Total Sum of Scores	Avg. Score per Hit	List of Scores	Medline	TF Description
S000002157	YBR089C-A	NHP6B	298	864.790127	2.90198029	[0.00026039439, 0.01	www.pubmed.com/1884262	High-mobility group (HMG)
S000006256	YPR052C	NHP6A	167	496.390153	2.97239613	[0.011299637, 0.0148	www.pubmed.com/1884262	High-mobility group (HMG)
S000002264	YDL106C	PHO2	89	434.072179	4.87721549	[0.89169651, 0.89169	www.pubmed.com/1911166	Homeobox transcription fa
S000005666	YOR140W	SFL1	105	342.958301	3.26626954	[0.0022765293, 0.004	www.pubmed.com/1884262	Transcriptional repressor a
S000000386	YBR182C	SMP1	125	302.509948	2.42007959	[0.061382908, 0.0733	www.pubmed.com/1884262	MADS-box transcription fa
S000000661	YCR065W	HCM1	81	300.987621	3.71589656	[0.02847782, 0.05670	www.pubmed.com/1911166	Forkhead transcription fact
S000002718	YDR310C	SUM1	48	220.615353	4.59615318	[0.5682655, 0.605034	www.pubmed.com/1911166	Transcriptional repressor th
S000001457	YIR018W	YAP5	43	184.145524	4.28245404	[0.49932978, 0.72623	www.pubmed.com/1652220	Basic leucine zipper (bZIP) i
S000004003	YLR013W	GAT3	77	180.992102	2.35054678	[0.067281015, 0.1040	www.pubmed.com/1911166	Protein containing GATA fa
S000004395	YLR403W	SFP1	70	165.241045	2.36058636	[0.0096045556, 0.076	www.pubmed.com/1884262	Regulates transcription of

Figure 4. Example output of the gene Msn2.

This .csv table shares the same format as its TF counterpart, except that the "Medline" column has the URLs to the reference which distinguished the PFM of the corresponding TF.

	Gene Feature Name	Gene Common Name	HMRA2 Number of Hits	HMRA2 Total Sum of Scores	DOT6 Number of Hits	DOT6 Total Sum of Scores	Gene Description
S000002159	YDL001W	RMD1	9	34.7110135	9	25.3192403	Cytoplasmic protein re
S000002158	YBR162W-A	YSY6	11	44.0045634	7	20.2708123	Protein of unknown fu
S000002153	YBL113C		6	18.9083203	13	50.1085678	Helicase-like protein er
S000002152	YBL112C		8	29.4250292	11	40.0281643	Putative protein of unk
S000002151	YBL111C		14	54.1705331	4	9.84809047	Helicase-like protein er
S000002150	YBL109W		13	49.6492261	5	11.5202028	Dubious open reading
S000002157	YBR089C-A	NHP6B	11	51.7601385	9	33.0818248	High-mobility group (H
S000002156	YBR084C-A	RPL19A	7	37.3647068	9	42.8762565	Ribosomal 60S subunit
S000002485	YDR078C	SHU2	11	48.9700713	10	31.5115438	Component of Shu com
S000002484	YDR077W	SED1	11	54.6789939	9	21.7969728	Major stress-induced s
S000002487	YDR080W	VPS41	12	50.2321153	7	19.4229524	Subunit of the HOPS er
S000002486	YDR079W	PET100	19	86.9731214	17	47.0180301	Chaperone that facilita
S000002481	YDR074W	TPS2	10	46.0526587	9	23.4124545	Phosphatase subunit o

Figure 5. Example output of shared gene targets between TFs HMRA2 and DOT6.

Unlike the .csv tables for individual TFs or genes, the .csv table representing shared gene targets or common TFs is not ranked by any column.

3.2 Individual TF and Gene Histograms

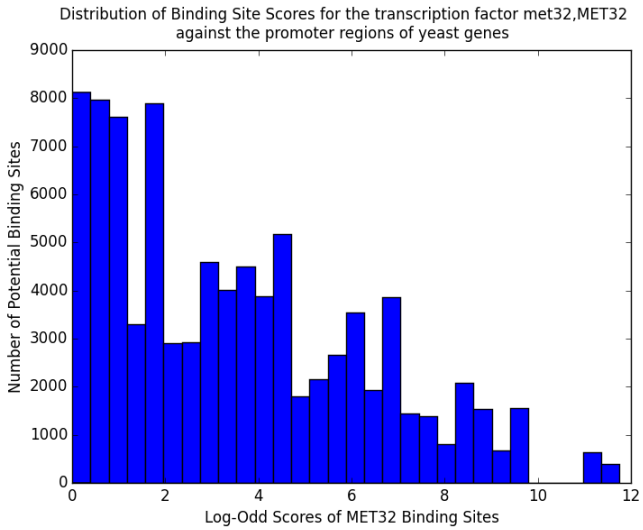

Figure 6. Histogram of TF Met32.

The distribution of the log-odd scores of Met32's potential TFBSs with Yeastract's promoter sequences indicates some clusters of potential TFBSs with similar scores (e.g. around 1.8, 4.2, and 6.2). The .csv table of Met32 can then be processed in order to isolate the potential TFBSs and associated promoter locations in order to see what patterns emerge.

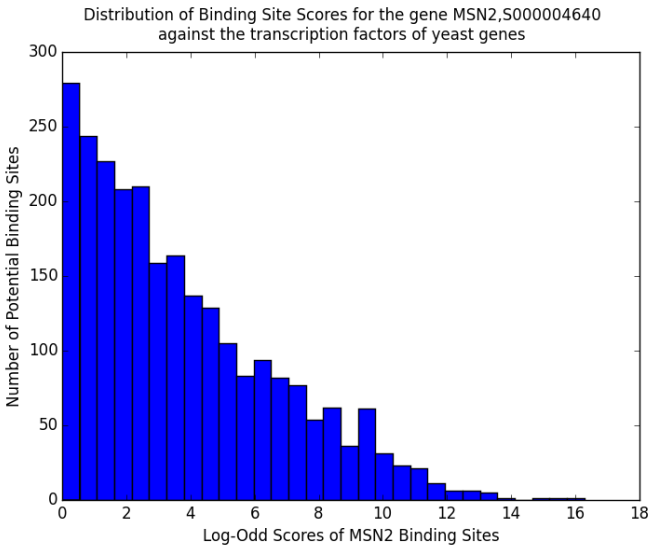

Figure 7. Histogram of gene Msn2.

The distribution of the log-odd scores of Msn2's potential TFBSs with JASPAR's PWMs does not appear to reveal any clear trends other than most of the potential TFBSs have a low log-odd score. Thus, further analysis of the gene Msn2 might benefit from an increase of the threshold in order to reveal new trends.

Discussion

4.1 Critical Needs of Improvement

While Brongus does not innovate on any of the tools currently at use in the field of TFBS prediction, the format and design of its analyses may offer a new perspective for outlining the interactions between one or more types of TFs and their cognate TFBSs. Given how dynamic homo- and heterosynergistic transcriptional regulation can become, an equally dynamic computational tool is required to account for the diverse factors that cooperatively induce transcriptional functionality in one or more genes.

Although there is certainly potential for Brongus to be developed further in aiding the characterization of synergistic transcriptional programs, it must first address the intrinsic faults of using a PWM model to locate putative TFBSs in a sequence, integrate more structural and locational information, and also update its default sets of genetic sequences and TF motifs.

Brongus currently determines potential TFBSs by scanning a PWM over each position in a sequence and calculating the log-odds probability score over the motif's length. However, this model of calculation is fraught with error– the relatively short motif length inevitably finds many potential TFBSs, the vast majority of which have no biological relevance. This high rate of sensitivity but terrible rate of accuracy is exacerbated by the model's latent assumptions that the binding of a TF to its cognate TFBS is independent and not affected by cofactors, interdependencies between nucleotides within the TFBS, competition from nucleosomes, or the flanking nucleotides. Another major fault in this model is the PWM itself. Unless PWM is generated from an abundance of high-quality experimental data, then it could be subject to inaccurate modeling of the binding preferences of its TF's motif. Furthermore, PWMs are assumed to have a fixed spatial binding, so they only bind to TFBSs of a specific length, which has been shown to be false for some TFs (Wassermann et al., 2004; Mathelier et al., 2013).

The overabundance of false TFBSs is compounded by the current default assembly of genetic sequences used by Brongus. Although Yeastract's promoter_sequences.fasta universally sets the promoter regions as 1000 base pairs in length, the YPA database indicates that the length of each promoter region is not regular, and can range from 300 to 1500 base pairs. This introduces the possibility of transcriptionally non-relevant sequences being analyzed for TFBSs, which slows down the speed of analysis, and skews our data even more.

Arguably the most major flaw with the current Brongus code is its complete exclusion of locational and structural data. Once a log-odds probability of a potential TFBS is calculated, the score is recorded, but not the position that yielded it. Given the current plethora of evidence that emphasizes the influence of nucleosome positioning, flanking nucleotides, rigid DNA structure, TSS positions, and the presence/absence of TATA boxes on the locations and transcriptional functionality of TFBSs, integrating even a modicum of this information would increase the rate of accuracy (Tirosh et al., 2007; Dror et al., 2016). However, the pertinent information is either not accessible for downloading, or whose integration into the code requires a

higher level of technical expertise that is currently unavailable. And although using YPA's resources to supplement the lack of structural information is a compelling notion, it is in desperate need of more regulatory information, as it was only last updated in 2011 (Chang et al., 2011).

And in addition to integrating more structural information and utilizing a default set of genetic sequences with better defined promoter regions, Brongus would also greatly benefit from updating from the current PWM model to the Transcription Factor Flexible Model (TFFM), which can greatly improve the accuracy of TFBS prediction by using experimental data and Bayesian Hidden Markov Models. This modeling system incorporates information from ChIP-seq data in order to generate flexible motif models that, unlike the PWM predecessor, can account for non-fixed motif lengths, dinucleotide dependencies, and accurately calculate the probability score of its occupancy for a given DNA region. Although JASPAR has made 130 TFFMs available since its last update, none of these TFFMs are designated for fungal TFs. However, as long as the pertinent ChiP-Seq data can be obtained, TFFMs can be generated with such data using a web-based app (Mathelier et al., 2015).

4.2 Future Improvements and Directions for Brongus

Assuming that the current issues with Brongus will be ameliorated, it is conceivable that its design and format could be adjusted in order to begin characterizing how the transcriptional functionality of certain genes is modulated by specific amplitude and/or temporal profiles of TFs. In addition to locating the TFBSs, calculating their respective binding affinities, and determining their TFFM occupancy scores, Brongus would also somehow start to delineate the temporal modulation of a genome's expression profile in response to a specific stimulus, perhaps by the integration of data that measured changes in activity for nuclear phosphatases, histone variants (e.g. H2A.Z) or nucleosome remodeling complexes. Then the mechanism underlying this and other similar *cis*-regulatory modules could be elucidated.

Acknowledgements

I would like to express my deepest, sincerest gratitude to my supervisors Dr. Julian Pietch and Dr. Peter Swain, as well as my bioinformatics professor Dr. Martin Jones, for the their unyielding support, assistance, and feedback. This project would not have been feasible without their contributions.

References

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X. and Kuznetsov, H., 2009. Diversity and complexity in DNA recognition by transcription factors. Science, 324(5935), pp.1720-1723.

Beskow, A. and Wright, A.P., 2006. Comparative analysis of regulatory transcription factors in Schizosaccharomyces pombe and budding yeasts. Yeast, 23(13), pp.929-935.

Borneman, A.R., Zhang, Z.D., Rozowsky, J., Seringhaus, M.R., Gerstein, M. and Snyder, M., 2007. Transcription factor binding site identification in yeast: a comparison of high-density oligonucleotide and PCR-based microarray platforms. Functional & integrative genomics, 7(4), pp.335-345.

de Boer, C.G. and Hughes, T.R., 2011. YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. Nucleic acids research, p.gkr993.

Brivanlou, A.H. and Darnell, J.E., 2002. Signal transduction and the control of gene expression. Science, 295(5556), pp.813-818.

Chang, D.T.H., Huang, C.Y., Wu, C.Y. and Wu, W.S., 2011. YPA: an integrated repository of promoter features in Saccharomyces cerevisiae. Nucleic acids research, 39(suppl 1), pp.D647-D652.

Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R. and Fisk, D.G., 2011. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic acids research, p.gkr1029.

Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. and de Hoon, M.J., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), pp.1422-1423.

Costanzo, M.C., Engel, S.R., Wong, E.D., Lloyd, P., Karra, K., Chan, E.T., Weng, S., Paskov, K.M., Roe, G.R., Binkley, G. and Hitz, B.C., 2014.

Saccharomyces genome database provides new regulation data. Nucleic acids research, 42(D1), pp.D717-D725.

Dalal, C.K., Cai, L., Lin, Y., Rahbar, K. and Elowitz, M.B., 2014. Pulsatile dynamics in the yeast proteome. Current Biology, 24(18), pp.2189-2194.

Davis, T.L., 2015. Package 'argparse'.

Dror, I., Rohs, R. and Mandel‐Gutfreund, Y., 2016. How motif environment influences transcription factor search dynamics: Finding a needle in a haystack. BioEssays.

Dervan, P.B., 1986. Design of sequence-specific DNA-binding molecules. Science, 232(4749), pp.464-471.

Farkas, I.J., Wu, C., Chennubhotla, C., Bahar, I. and Oltvai, Z.N., 2006. Topological basis of signal integration in the transcriptional-regulatory network of the yeast, Saccharomyces cerevisiae. BMC bioinformatics, 7(1), p.1.

Gao, Z., Zhao, R. and Ruan, J., 2013. A genome-wide cis-regulatory element discovery method based on promoter sequences and gene co-expression networks. BMC genomics, 14(1), p.1.

Goldschmidt, Y., Yurkovsky, E., Reif, A., Rosner, R., Akiva, A. and Nachman, I., 2015. Control of relative timing and stoichiometry by a master regulator. PloS one, 10(5), p.e0127339.

Guigo, R. (2003) *BIOINFORMÀTICA [online].* Barcelona, Universitat Pompeu Fabra. Available from: http://bioinformatica.upf.edu/T12/MakeProfile.html [14/08/2016].

Guthrie, C. and Fink, G.R. eds., 2002. Guide to yeast genetics and molecular and cell Biology: Part C (Vol. 350 and 351). Gulf Professional Publishing.

Hansen, A.S. and O'Shea, E.K., 2013. Promoter decoding of transcription factor dynamics involves a trade‐off between noise and control of gene expression. Molecular systems biology, 9(1), p.704.

Hansen, A.S. and O'Shea, E.K., 2015. Limits on information transduction through amplitude and frequency regulation of transcription factor activity. Elife, 4, p.e06559.

Hansen, A.S. and O'Shea, E.K., 2016. Encoding four gene expression programs in the activation dynamics of a single transcription factor. Current Biology, 26(7), pp.R269-R271.

Hu, J. and Zhang, J., 2010, April. Co-occurrence of core of binding sites for transcription factors in intronic region of Saccharomyces cerevisiae

ribosomal protein genes. In Bioinformatics and Biomedical Technology (ICBBT), 2010 International Conference on (pp. 88-91). IEEE.

Lam, F.H., Steger, D.J. and O'Shea, E.K., 2008. Chromatin decouples promoter threshold from dynamic range. Nature, 453(7192), pp.246-250.

Lee, W., Tillo, D., Bray, N., Morse, R.H., Davis, R.W., Hughes, T.R. and Nislow, C., 2007. A high-resolution atlas of nucleosome occupancy in yeast. Nature genetics, 39(10), pp.1235-1244.

Li, Q. and Johnston, S.A., 2001. Are all DNA binding and transcription regulation by an activator physiologically relevant?. Molecular and cellular biology, 21(7), pp.2467-2474.

Lin, Y., Sohn, C.H., Dalal, C.K., Cai, L. and Elowitz, M.B., 2015. Combinatorial gene regulation by modulation of relative pulse timing. Nature.

Lubliner, S., Keren, L. and Segal, E., 2013. Sequence features of yeast and human core promoters that are predictive of maximal promoter activity. Nucleic acids research, p.gkt256.

Mathelier, A. and Wasserman, W.W., 2013. The next generation of transcription factor binding site prediction. PLoS Comput Biol, 9(9), p.e1003214.

Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C.Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R. and Zhang, A.W., 2015. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic acids research, p.gkv1176.

Meyer, T. and Vinkemeier, U., 2004. Nucleocytoplasmic shuttling of STAT transcription factors. European Journal of Biochemistry, 271(23‐24), pp.4606-4612.

Moll, T., Tebb, G., Surana, U., Robitsch, H. and Nasmyth, K., 1991. The role of phosphorylation and the CDC28 protein kinase in cell cycle-regulated nuclear import of the S. cerevisiae transcription factor SW15. Cell, 66(4), pp.743-758.

Morse, R.H., 2007. Transcription factor access to promoter elements. Journal of cellular biochemistry, 102(3), pp.560-570.

Pachkov, M., Erb, I., Molina, N. and Van Nimwegen, E., 2007. SwissRegulon: a database of genome-wide annotations of regulatory sites. Nucleic acids research, 35(suppl 1), pp.D127-D131.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A. and Segal, E., 2012. Inferring gene regulatory logic from high-throughput measurements of

thousands of systematically designed promoters. Nature biotechnology, 30(6), pp.521-530.

Stewart, A.J., Hannenhalli, S. and Plotkin, J.B., 2012. Why transcription factor binding sites are ten nucleotides long. Genetics, 192(3), pp.973-985.

Stormo, G.D., 2013. Modeling the specificity of protein-DNA interactions. Quantitative biology, 1(2), pp.115-130

Tanay, A., 2006. Extensive low-affinity transcriptional interactions in the yeast genome. Genome research, 16(8), pp.962-972.

Teixeira, M.C., Monteiro, P.T., Guerreiro, J.F., Gonçalves, J.P., Mira, N.P., dos Santos, S.C., Cabrito, T.R., Palma, M., Costa, C., Francisco, A.P. and Madeira, S.C., 2013. The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae. Nucleic acids research, p.gkt1015.

Tkach, J.M., Yimit, A., Lee, A.Y., Riffle, M., Costanzo, M., Jaschob, D., Hendry, J.A., Ou, J., Moffat, J., Boone, C. and Davis, T.N., 2012. Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. Nature cell biology, 14(9), pp.966-976.

Todeschini, A.L., Georges, A. and Veitia, R.A., 2014. Transcription factors: specific DNA binding and specific gene regulation. Trends in Genetics, 30(6), pp.211-219.

Vorontsov, I.E., Kulakovskiy, I.V. and Makeev, V.J., 2013. Jaccard index based similarity measure to compare transcription factor binding site models. Algorithms for Molecular Biology, 8(1), p.1.

Wasserman, W.W. and Sandelin, A., 2004. Applied bioinformatics for the identification of regulatory elements. Nature Reviews Genetics, 5(4), pp.276-287.

Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V. and Romano, L.A., 2003. The evolution of transcriptional regulation in eukaryotes. Molecular biology and evolution, 20(9), pp.1377-1419.

Zhang, Z. and Dietrich, F.S., 2005. Mapping of transcription start sites in Saccharomyces cerevisiae using 5′ SAGE. Nucleic acids research, 33(9), pp.2838-2851.