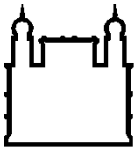


Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Doutorado em Biologia Computacional e Sistemas

Determinantes e forças seletivas na evolução das proteínas

Luis Fernando Encinas Ponce

Tese apresentada à Coordenação do Curso de Doutorado em Biologia Computacional e Sistemas como requisito parcial para obtenção do título de Doutor em Ciências

Orientador: Dr. Antonio Basílio de Miranda

Rio de Janeiro

2014

Ficha catalográfica elaborada pela

Biblioteca de Ciências Biomédicas/ ICICT / FIOCRUZ - RJ

E56 Encinas Ponce, Luis Fernando

Determinantes e forças seletivas na evolução das proteínas / Luis Fernando Encinas Ponce. – Rio de Janeiro, 2014.

xiii, 132 f.: il. ; 30 cm.

Tese (Doutorado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2014.

Bibliografia: f. 80-86

1. Evolução de proteínas. 2. Mineração de dados. 3. Sistemas biológicos. I. Título.

CDD 572.6

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Doutorado em Biologia Computacional e Sistemas

Determinantes e forças seletivas na evolução das proteínas

Luis Fernando Encinas Ponce

ORIENTADOR: Dr. Antonio Basílio de Miranda

Banca examinadora

Dr. Oswaldo Gonçalves Cruz (Presidente)

Dra. Renata Schama Lellis

Dr. Marcos Catanho de Souza

Dr. Alberto Rivera Dávila

Dr. Gonzalo Bello Betancour

Rio de Janeiro, Março de 2014

A meus tesouros Eugene e Marcia

Agradecimentos

Agradeço primeiramente aos Professores, colegas e pessoal administrativo da Pós-Graduação em Biologia Computacional e Sistemas por a instrução, incentivo e ajuda que recebi nos últimos quatro anos.

Ao meu orientador, Prof. Antonio Basílio pela guia, amizade e apoio para fazer esta tese uma realidade. Valeu chefe!!

A minha esposa Marcia e meu filho Eugene por tanto amor, tanto apoio e por serem eles a motivação pela superação. Amo muito!!

Aos meus pais, Lourdes, Raúl e Fernando, pelo exemplo, incentivo e refugio de sempre. A saudade é grande mas o amor é imenso!

Aos meus amigos, Leandro, Monete, Michel, Marcio, Lalá. Queridíssimas pessoas que pude conhecer nesta desafiante empreitada. Tamo junto!

A todo o pessoal dos laboratórios de Biologia Computacional e Sistemas e Bioinformática e Genômica Funcional.

Finalmente, a todas as pessoas que direta ou indiretamente me apoiaram para concretizar a finalização deste trabalho.

Muito obrigado!!

Lista de figuras

Lista de tabelas

Lista de anexos

TABELA DE CONTEÚDO

Capítulo II Análises de custo e benefício da regulação cinética traducional

"Biology has changed dramatically, becoming one of the most mathematics- and data-intensive of all the sciences. If its culture does not fully embrace the intellectual challenge presented by its own data models, it will forever fall short of its potential."

Tony Berno, Nature, Vol. 499, 7456 (2013)

RESUMO

A análise de grandes quantidades de dados aproveitando o poder computacional de ferramentas "open source" que estão disponíveis na internet é o que veio a conhecer-se como quarto paradigma da investigação científica[ξ].

Em muitas áreas do conhecimento como a Astronomia, a Física e Geologia, a experimentação, o desenvolvimento teórico e o poder computacional (os três primeiros paradigmas) têm dado lugar à análise rotineira de grandes quantidades de dados e o desenvolvimento de novos métodos, conceitos e teorias que permitam interpretar a informação gerada por novas tecnologias.

No campo da biologia, esta mudança nos paradigmas da investigação científica supõe um desafio na hora de encarar uma questão biológica; mas, em contrapartida, ela oferece a oportunidade de validar teorias clássicas e/ou testar hipóteses novas.

Precisamente neste contexto, a presente tese aborda duas questões pertinentes ao campo da biologia evolutiva: *Quais são os fatores que determinam a evolução de uma proteína?* e *Qual é a natureza da seleção cinética traducional?*. Estas perguntas são, em principio, relevantes no âmbito teórico; por outro lado, sua compreensão, implicações e perspectivas têm também espaço importante na área experimental.

A tese está estruturada da seguinte forma:

No **Capitulo um** se descreve uma combinação de análise de texto com outras técnicas de mineração de dados para identificar, classificar, integrar e modelar associações existentes entre caracteres genômicos que favorecem ou impedem a acumulação de substituções nucleotídicas ao nível das regiões codificadoras.

Nossa metodologia permitiu identificar características genômicas como a eficiência traduçional, a instabilidade estrutural e as regiões de baixa complexidade que em principio poderiam constituir determinantes da evolução das proteínas.

Construtos latentes como esquema de integração de dados biológicos mostraram que, em vez de considerar o nível de mRNA como o maior determinante da evolução das proteínas, outras variáveis relacionadas com a expressão de um gene podem ser igualmente importantes.

Finalmente, graças a um modelo de fatores Bayesiano, foi possível estimar os componentes de um sistema de tradução de proteínas identificado com a eficiência e adaptação da maquinaria celular.

No **Capitulo dois,** o controle cinético exercido pelos códons raros durante a tradução das proteínas é abordado com a ajuda de uma análise de custo-benefício que tenta identificar a natureza do que veio a denominar-se como seleção cinética traducional. Diferenças entre proteínas estáveis e instáveis apóiam permitiram identificar a ação da regulação cinética traducional sobre determinado grupos de genes.

Os padrões de substituções sinônimas encontrados nas proteínas instáveis permitiram estender nossa discussão apontando à existência de combinações de códons

num espaço genotípico determinado que assegure a conservação da estrutura terciária de uma proteína, mas, ao mesmo tempo procure a otimização da cinética da sua tradução.

ξ:Em uma série de conferências no ano de 2007, um investigador da Microsoft Research, James Gray (1944- 2012) apresentou um argumento no qual ele afirmava que o poder computacional disponível teria mudado para sempre a pratica da ciência.

O Dr. Gray chamou esta mudança como "O quarto paradigma da investigação cientifica". Sendo os três primeiros paradigmas o experimental, o teórico e o mais recente, o computacional; ele explicou este paradigma como a evolução de uma era na qual uma inundação de dados observacionais ameaçava inviabilizar os cientistas. A única maneira de lidar com ela, segundo ele, era uma nova geração de computação científica incluindo novas ferramentas para gerenciar, visualizar e analisar os dados.

ABSTRACT

In scientific discovery, three acknowledged paradigms are experimental, theoretical and computational. In the last ten years however, scientists have been overwhelmed with large amounts of data coming from high-throughput technologies that are analyzed taking advantage of computational power, the internet and open source data-analysis tools.

Late researcher of Microsoft, Dr. James Gray (1944-2012 *in absentia*) called this "the fourth paradigm of scientific research" and urged the need to acknowledge that making sense of data will turn routine in most areas of science.

For biologists and others involved in life sciences, this paradigm shift may address daunting challenges, however; in return, it offers the opportunity to examine old theories and test new hypothesis.

It is within this context that the thesis presented here tackles two fundamental problems of evolutionary biology: What are the constraints of protein evolution? and what is the underlying nature of the kinetic-translational selection?.

Although at first glance these questions might appear exclusively relevant for the theoretical field of evolutionary biology, we consider their implications for other areas such as biotechnology and clinical applications.

The thesis is organized as following:

In **Chapter one**, we present a combination of text analysis with other data mining techniques to identify, classify, integrate and model existing associations between genomic characters that favor or hinder the rate at which proteins evolve.

Our methodology allowed us to identify genomic features such as translational efficiency, structural instability and low-complexity regions that appear to constitute constraints of protein evolution.

Latent constructs were used as an alternative to integrate biological data and they showed that instead of using mRNA levels as primary determinants of protein evolution, other expression-related factors should be considered.

We devised a Bayesian factor model to estimate the components of a protein translation system identified with the efficiency and adaptation of the cellular machinery.

In **Chapter two**, we aboard the fine-tuning kinetic control of rare codons during protein translation in the context of a cost-benefit analysis devised to identify the action of recently proposed kinetic translational selective force.

The pattern of synonymous substitutions found in proteins classified as structurally unstable led us to extend our discussion to the existence of a determined genotypic space in which combinations of codons are “tested” in order to optimize the protein synthesis kinetics maintaining the tridimensional structure.

CAPÍTULO 1

MINERAÇÃO, INTEGRAÇÃO E MODELAGEM DE FATORES GENÔMICOS QUE DETERMINAM A EVOLUÇÃO DAS PROTEÍNAS

1. Introdução

As causas de variação nas taxas evolutivas das proteínas têm sido um tópico de interesse recorrente no campo da biologia evolutiva (Pál & Lercher, 2006; Lucas-Lledó & Lynch, 2009; Du *et al*., 2013). Diversas análises de genômica comparativa permitiram a identificação de fatores individuais, funcionais e estruturais, que favorecem ou dificultam a taxa em que as substituições se acumulam ao nível dos nucleotídeos (Vieira-Silva *et al*., 2011; Coulombe-Huntington & Xia, 2012; Chakraborty *et al*., 2010). Entre estes fatores, embora alguns exemplos contrários existam (Tirosh & Bakrai, 2008), o nível de expressão gênica foi indicado como o principal determinante da evolução das proteínas (Drummond *et al*., 2006; Gout*et al*., 2010).

O acesso a diferentes tipos de informação biológica confirmou a complexidade dos organismos como sistemas vivos (Berger *et al*., 2013) e mudou nosso entendimento sobre as margens fenotípicas nas quais a seleção pode operar (Koonin & Wolf, 2010). Portanto, à luz da crescente quantidade de dados experimentais, existe a necessidade de reexaminar os fatores que determinam as mudanças evolutivas e de integrar os dados relacionados para abordar o problema da evolução das proteínas a partir de uma perspectiva holística.

A integração de dados relacionados é particularmente proveitosa já que permite extrair o valor real de cada um dos conjuntos de dados; porém, para tornar essa integração viável e significativa, é necessária a aplicação de métodos computacionais avançados, acompanhados muitas vezes por métodos matemáticos e

estatísticos adequadamente sustentados numa estrutura teórica (Gopalacharyulu *et al*., 2005).

A mineração de dados como ciência aplicada é o processo, assistido por um computador, de analisar grandes quantidades de dados para descrevê-los e resumi-los em informação relevante (Besmail & Haoudi, 2005). Através de uma grande variedade de técnicas, a mineração de dados permite o reconhecimento de padrões que não são imediatamente evidentes e têm a flexibilidade de explicar os dados tanto ao nível individual como ao nível de sistemas (Rebholz-Schuhmann *et al.*, 2012).

No presente capítulo se apresenta uma metodologia combinada que, começando com análises de texto, coleta dados de variáveis genômicas que podem constituir-se em determinantes da evolução das proteínas. Métodos avançados de clusterização hierárquica e análises de fatores foram utilizados para explicar a estrutura do conjunto de dados a um nível mais elevado e, por último, um modelo de fatores Bayesiano foi testado para estimar os componentes do que seria um sistema de tradução de proteínas eficiente.

2. Referencial teórico

2.1 Forças que dirigem a evolução das espécies

A evolução de um organismo é um processo de acumulação de mudanças genéticas, resultado de uma variedade de mecanismos moleculares condicionados a vários níveis da organização biológica que são efetivadas pela ação individual ou conjunta de várias forças evolutivas num determinado fenótipo (Carey, 2003).

Assim, num contexto de tempo e hereditariedade, são basicamente as interações entre as forças evolutivas, os mecanismos de variabilidade genética e os condicionantes desta variabilidade, que determinam a historia evolutiva dos organismos e das espécies as quais pertencem.

Embora exista alguma disputa sobre a importância relativa de cada uma, é bem aceito que são quatro as principais forças que governam a evolução das espécies: a seleção natural, a deriva genética, as mutações e o fluxo gênico (Carey, 2003).

A seleção natural é a única força evolutiva que pode resultar na geração de caracteres adaptativos na procura pela harmonização entre um organismo e o meio ambiente, ou na eliminação de caracteres prejudiciais (Futuyma, 2009).

O efeito do acaso em populações pequenas é o que se conhece como deriva genética. É nestas populações que erros de amostragem se tornam mais evidentes e podem alterar as frequências dos alelos de uma geração a outra (Graur & Li, 2000).

As mutações são a maior fonte de variação genética dentro de uma população e embora a maior parte delas possam ser neutras (com nenhum efeito na aptidão, em inglês, *fitness*), outras podem ter um pequeno efeito positivo e são essas variantes as

que constituem a matéria-prima da evolução adaptativa (Sniegowski & Lenski, 1995).

A força da migração ou fluxo gênico tem efeitos na variabilidade genética que são opostos aos causados pela deriva genética. A migração limita a divergência genética das populações e desta forma impede o processo de especiação (Lenormand, 2002).

2.2 A seleção natural

A seleção natural é definida como a reprodução diferencial de um organismo em função de caracteres herdáveis que influem na adaptação ao meio ambiente.

O conceito de seleção natural é fundamental para a teoria de Charles Darwin e constitui a pedra angular de muitos estudos no campo da evolução. Como já foi referido anteriormente, a seleção natural é o único mecanismo de evolução adaptativa e é preciso pensar nela mais como um processo gradual que como uma força guia (Futuyma, 2009).

A seleção natural pode manter ou eliminar a variação genética dependendo de como ela age. Quando alelos deletérios são eliminados, ou quando impede que um alelo se fixe na população, a seleção natural diminui a variação genética. Quando heterozigotos de alguma forma são mais adaptados que qualquer um dos homozigotos, a seleção natural mantém a variação genética (Bulmer, 1971).

Dependendo então de como ela age, a seleção natural pode levar uma população numa variedade de direções. Assim, a seleção disruptiva serve para incrementar a frequência de fenótipos raros e diminuir a frequência daqueles comuns. A seleção direcional pode resultar numa mudança na frequência de um ou mais caracteres em uma direção particular. E a seleção estabilizadora atua em contra dos fenótipos extremos e favorece os fenótipos mais comuns dentro da população (Brodie *et al*., 1995).

A seleção natural não tem nenhuma antevisão ou projeto. Ela apenas permite aos organismos a se adaptarem ao seu ambiente atual. Estruturas ou comportamentos não evoluem para uma utilidade futura. Um organismo está adaptado para seu ambiente em cada respectivo estágio de sua evolução. Com as mudanças ambientais, novos caracteres podem ser selecionados favoravelmente.

2.3 Mecanismos de variabilidade genética

Para que a evolução possa acontecer, mecanismos que criem variação genética devem existir.

Durante o processo de replicação do Ácido Desoxirribonucléico (ADN) uma cópia exata da fita molde é criada. No entanto, um ou vários erros na incorporação do nucleotídeo correto na replicação ou mesmo durante o processo de reparo existem e estes são conhecidos como mutações (Pray, 2008).

Assim, dependendo do tipo de mudança causada ao nível do DNA as mutações podem ser classificadas em:

2.3.1 Mutações substitutivas: Divididas entre transições e transversões, uma **transição** ocorre quando existe uma substituição de uma base nitrogenada por outra do mesmo grupo (uma purina por outra purina, ou uma pirimidina por outra pirimidina) enquanto uma **transversão** ocorre quando a base nitrogenada é substituída por uma do outro grupo (uma purina por uma pirimidina ou vice-versa) (Garduño *et al.*, 1977).

Devido à estrutura do código genético, as mutações substitutivas que ocorrem nas regiões codificadoras de proteínas podem ser classificadas em **não-sinônimas** se elas causarem a substituição do aminoácido especificado por algum outro, e **sinônimas** se a substituição não tem efeito algum na seqüência de aminoácidos resultante (Graur & Li, 2000).

2.3.2 Recombinação: Constitui o intercâmbio de uma seqüência por outra e pode ser classificada em **recombinação recíproca** quando existe um intercâmbio equivalente de sequências homólogas entre cromossomas homólogos e **recombinação não-recíproca** quando o intercâmbio envolve a substituição não equilibrada de uma sequência por outra (Sherman & Roman, 1963).

Enquanto a recombinação recíproca produz novas combinações de sequências adjacentes reunindo ambas as variantes envolvidas no evento de recombinação, a recombinação não-recíproca resulta na perda de uma das sequências envolvidas na recombinação; tem sido sugerido que, junto com a substituição nucleotídica, a recombinação homóloga (especialmente a recombinação recíproca) são os maiores geradores da variabilidade genética (Lercher & Hurst, 2002).

2.3.3 Deleções e Inserções: Conhecidos coletivamente como ***indels***, inserções e deleções podem ocorrer por vários mecanismos. Quando duas sequências são comparadas entre si, é muito difícil determinar se o que ocorreu foi uma deleção em uma delas ou uma inserção na outra. Em geral, o comprimento dos *indels* exibe uma distribuição de freqüência bimodal, com *indels* curtos de vinte a trinta nucleotídeos principalmente causados por erros na replicação, e inserções ou deleções longas resultantes de mecanismos tais como recombinação sítio-específica, transposição, transferência horizontal ou *crossing-over* desigual (revisado em Mullaney *et al.*, 2010).

Em sequências codificadoras, um *indel* tem capacidade de alterar a fase de leitura na região posterior ao *indel* se ele não ocorrerem um múltiplo de três, podendo desta forma não só introduzir várias mudanças na incorporação de aminoácidos errados, como também provocar a terminação prematura da leitura resultando assim numa proteína de menor comprimento (Garcia-Diaz & Kunkel, 2006).

2.3.4 Inversões: Inversões são tipos de rearranjos de DNA que podem ocorrer como resultado de uma incisão e posterior reunião cromossômica ou como consequência de um *crossing-over* entre dois segmentos homólogos que estão orientados em direções opostas. Em geral as inversões envolvem segmentos de DNA muito compridos de centenas ou milhares de nucleotídeos (Graur & Li, 2000).

2.4 Taxas de substituição nucleotídica

Como dito anteriormente, as mutações são a fonte principal de novidade genética; por conseguinte, determinar a taxa à qual surgem novas mutações é uma questão central em genética (Nachman, 2004). Comumente, estas taxas são medidas pelo número de substituições entre duas sequências codificadoras, e vários métodos têm sido desenvolvidos para estimar as taxas de substituição sinônimas (Ks) e não-sinônimas (Ka) (Tzeng*et et al*., 2004). Estas taxas constituem a abordagem mais direta para quantificar a importância relativa da seleção e deriva genética e para inferir o tempo de eventos evolutivos importantes, como especiação (Nachman & Crowell, 2000).

Comparações genômicas extensas permitiram observar que as taxas evolutivas entre proteínas variam por várias ordens de magnitude, e as causas desta variação foram sempre um tema de muita discussão (Pálet *et al.*, 2006).

2.5 Restritores seletivos na variação das taxas substitutivas

A seleção natural atua através de um mecanismo conhecido como restrição seletiva. Quando um gene, uma via bioquímica, ou um caráter fenotípico é "restrito seletivamente", ele é mantido ao longo do período evolutivo (Arnold, 1992).

São muitos os níveis nos quais a restrição seletiva pode atuar; por exemplo, uma via bioquímica poderia ser tão fundamental para a capacidade de sobrevivência de um organismo que qualquer alteração nesta via poderia ter efeitos letais. Uma única mutação em um gene que codifica uma proteína essencial poderia alterar a estrutura da proteína e torná-la não funcional (Wang *et al.*, 2004).

Desta forma, a seleção natural e a restrição seletiva são dois importantes paradigmas para entender a evolução. Eles não são toda a história, mas eles nos

ajudam a entender como a evolução produz mudanças, mas também propaga as semelhanças.

Tradicionalmente, a expressão gênica, a estrutura tridimensional e a função foram consideradas como principais restritores ou determinantes da evolução das proteínas. É notável que muitos trabalhos coincidem ao afirmar que o nível de expressão gênica é o fator mais importante, explicando quase 50% da variação da taxa de evolução das proteínas (Drummond *et al.*, 2006), e que a disponibilidade da informação do genoma derivado a partir das sequências de nucleotídeos completas e perfis de expressão permitiram observar que, em geral, genes altamente expressos evoluem lentamente enquanto os genes que evoluem rapidamente tendem a se expressar em níveis baixos (Subramanian & Kumar, 2004).

Devido à necessidade de formar e manter o local ativo definitivo, o que provavelmente exerce uma forte pressão seletiva para que uma proteína adote um enovelamento estável e conservado, a estrutura das proteínas tem sido geralmente considerada como o "registro fóssil" da evolução molecular (Andreeva & Murzim, 2006). No entanto, à medida que mais estruturas de proteínas tornam-se disponíveis e mais projetos de genômica estrutural geram informação nova e inédita, uma importante questão biológica é: Como as propriedades físicas de um sistema influenciam a sua capacidade para evoluir? (Bloom *et al.*, 2006). Todas as limitações relacionadas com a manutenção da estrutura terciária são eventualmente funcionais. Muitas funções são mediadas através de interações quaternárias de proteínas com outras macromoléculas, assim, em termos de importância, a pressão por manter a atividade de uma proteína será maior quanto mais essencial for essa proteína para assegurar a sobrevivência do organismo (Worth *et al.*, 2009).

2.6 Os desafios na era pós-genômica: A complexidade biológica e a integração dos dados

O impacto de projetos genômicos se traduz não só em uma maior quantidade de informações de sequência. A disponibilidade dos diferentes tipos de dados experimentais de alta vazão reafirmou a complexidade de organismos como sistemas vivos e, por conseguinte, para obter uma compreensão integrada de formas de vida em vários níveis, esta deve estar intimamente ligada a um componente evolutivo (Koonin & Wolf, 2006).

Este componente evolutivo que se concentra basicamente na interação entre genótipo e fenótipo foca-se na identificação e correlação de variáveis genômicas que determinam restrições seletivas, e em analisar como as mudanças em um nível refletem sobre a evolução em outro nível (Koonin & Wolf, 2006).

Até agora, diferentes fatores com relativa influência nas taxas evolutivas das proteínas têm sido identificados. Variações genômicas nas taxas mutacionais, nas taxas de recombinação, nos níveis de expressão, na dispensabilidade, nas interações e ainda outras relacionadas com as propriedades individuais das proteínas (revisado em Pál *et al.*, 2006), em certa medida contribuem para dificultar ou favorecer a taxa com a qual as substituições se acumulam ao nível de nucleotídeos.

Infelizmente, na tentativa de explicar a relação entre a evolução do genoma e o fenótipo dos organismos, a falta, imprecisões e distorções nos dados analisados são discutíveis e a inadequação dos modelos teóricos existentes também representa uma grande limitação (Koonin, 2005).

Assim, há necessidade de cenários alternativos que permitam testar hipóteses clássicas e o estabelecimento de novas teorias e novas formas de estudar os processos evolutivos (Medina, 2005).

Uma vez que hoje em dia os dados de alta vazão são digitalmente armazenados em uma ampla variedade de formatos (bases de dados), novos métodos computacionais são continuamente desenvolvidos para a mineração e análise de tais dados (Lacroix, 2002). O valor de cada conjunto de dados, no entanto, só pode ser apreciado, se eles são combinados ou integrados em uma única estrutura (Almeida *et al.*, 2006). Desta forma, a integração de dados heterogêneos é um grande objetivo, mas enorme desafio que pode ser abordado de duas maneiras diferentes: lidando com arquiteturas de bancos de dados, ferramentas de software e ontologias. A integração de banco de dados persegue a complementação e compreensibilidade das informações obtidas a partir da web (Gopalacharyulu *et al.*, 2005) e pode ser imaginado com uma "integração física".

De uma maneira diferente, a "pesquisa baseada em modelo" foca-se na integração de dados relacionados, apoiando-se em diferentes áreas da ciência, como a matemática, física, ciência da computação e estatística, para simular o comportamento de um sistema de modo a compreender os seus mecanismos biológicos (Yao, 2002).

2.7 Disponibilidade, armazenamento e organização da informação biológica

O crescimento acelerado do volume e tipos de dados na área da Biologia se deve ao desenvolvimento de técnicas de laboratório que permitem a coleta dos mesmos através de equipamentos sofisticados. Esta imensa quantidade de dados deve ser organizada de maneira acessível a modo de facilitar sua posterior análise;

consequentemente, a construção de bancos de dados para o armazenamento de informação em sequências de DNA, genomas completos, estrutura das proteínas, expressão gênica e outros da era genômica, tem sido, e ainda continua sendo uma área fundamental e de muito estudo no campo da Bioinformática (Baxevanis, 2011).

Os diferentes tipos de informação armazenada e a importância dos bancos de dados no desenvolvimento da pesquisa na área da biologia se vêm refletidos no incremento no número de bancos de dados biológicos listados na edição anual da coleção de bancos de dados do Journal of Nucleic Acid Reserach (NAR); 1512 bancos entre os anos de 1999 e 2013 (Fernandez-Suarez & Galperim, 2012). Este número porém, poderia ser maior se os bancos de dados criados antes de 1999 fossem contabilizados.

Basicamente, existem três tipos de bancos de dados:

- **Bancos de dados primários:** Nos quais os dados armazenados provêm diretamente de algum método de laboratório, por tanto o conteúdo é controlado pelo pesquisador que submete os dados. Bancos de dados primários são o GenBank, ENA, DDBJ, GEO e PDB.

O GenBank (http://www.ncbi.nlm.nih.gov/genbank/) é o principal banco de dados do NCBI e armazena todas as sequências disponíveis publicamente de DNA (de sequências pequenas a genomas inteiros), RNA e proteínas. Outros dois bancos de dados similares estão localizados na Europa (ENA/EBI) (http://www.ebi.ac.uk/ena/) e no Japão (DDBJ) (http://www.ddbj.nig.ac.jp/) e eles trocam dados em um intervalo de 24 horas.

O GEO (http://www.ncbi.nlm.nih.gov/geo/) foi criado para armazenar dados de expressão gênica e de hibridação de genomas enquanto o PDB

(http://www.rcsb.org/pdb/home/home.do) é um banco de dados de estruturas de proteínas e ácidos nucléicos determinados experimentalmente através da difração de raios X ou da ressonância magnética nuclear.

- **Bancos de dados secundários:** Também chamados bancos de dados derivados, estes são construídos em base a padrões encontrados na análise dos bancos de dados primários e são os curadores os responsáveis pela informação armazenada. Alguns exemplos de bancos de dados secundários são: RefSeq, Pfam, COGs, CDD, UniprotKB/Swiss-Prot, InterPro.

O SWISS-PROT foi criado em 1986 e atualmente é mantido pelo Swiss Institute of Bioinformatics (SIB) e o EMBL/EBI (http://www.ebi.ac.uk/uniprot). Este banco mantém um alto nível de anotações, como a descrição e função de proteínas, estrutura dos seus domínios e modificações pós-traducionais entre outros.

Muitas proteínas são construídas a partir de domínios em uma arquitetura modular; por tanto, o estudo de famílias de proteínas é melhor englobado como um estudo de famílias de domínios de proteínas. Prodom (http://prodom.prabi.fr/prodom/current/html/home.php) e CDD (http://www.ncbi.nlm.nih.gov/cdd/) são bancos de dados de sequências de domínios de proteínas criados automaticamente a partir de bancos de dados primários.

O InterPro (http://www.ebi.ac.uk/interpro/) é um banco de dados de assinaturas, capacitado para identificar relacionamentos distantes entre novas seqüências, conseguindo, assim, inferir funções protéicas. Como uma base integrada de documentação de famílias de proteínas, domínios e regiões funcionais, o InterPro integra os esforços do PROSITE (http://prosite.expasy.org/), do PRINTS (http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php), do Pfam

(http://pfam.sanger.ac.uk/) e do ProDom. Cada entrada do InterPro inclui uma descrição funcional, uma anotação e referências da literatura, além de *links* para os bancos de dados importantes.

Nesta classificação, ainda é possível distinguir outro tipo de **bancos de dados** os quais podem ser chamados de "**agregados**" ou **especializados**. Entre estes temos, por exemplo, os bancos de dados bibliográficos como o PUBMED ou MEDLINE (http://www.ncbi.nlm.nih.gov/pubmed), bancos de dados de metabolismo como o KEGG (http://www.genome.jp/kegg/) e bancos de dados descritivos como o Gene Ontology (http://www.geneontology.org/) cuja organização hierárquica tenta estandardizar a representação de um produto gênico a diferentes níveis.

2.8 Mineração de Dados

A Mineração de Dados ou "*data mining*" em inglês é um termo genérico para uma variedade de técnicas analíticas cujo objetivo principal é a busca de padrões ocultos dentro de grandes conjuntos de dados (Oliveira & da Silva, 2009). Estas técnicas têm sido restritas a campos tais como Psicologia e Sociologia por muito tempo; no entanto, o crescimento explosivo da internet, as grandes quantidades de dados e o processamento computacional contribuíram para seu ressurgimento e hoje estas técnicas estão presentes em todos os campos da ciência, incluindo a genômica e a proteômica (Bensmail & Haoudi, 2005).

Por ser uma área considerada multidisciplinar, as definições acerca da Mineração de Dados variam com o campo de atuação dos autores. Uma definição abrangente: "Mineração de Dados é um passo no processo de descoberta de conhecimento que consiste na realização da análise dos dados e na aplicação de

algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados" (Fayyad *et al.*, 1996).

A Mineração de Dados é comumente classificada pela sua capacidade em realizar determinadas tarefas. As mais comuns são:

- **Descrição**: Tarefa para descrever os padrões e tendências revelados pelos dados;

- **Classificação**: Visando identificar a qual classe um determinado registro pertence;

- **Estimação** ou **Regressão**: A estimação é similar à classificação, porém é usada quando o registro é identificado por um valor numérico e não um categórico. Assim, pode-se estimar o valor de uma determinada variável analisando-se os valores das demais;

- **Agrupamento** ou **Clusterização**: A tarefa de agrupamento visa identificar e aproximar os registros similares. Um agrupamento (ou *cluster*) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Esta tarefa difere da classificação pois não necessita que os registros sejam previamente categorizados (aprendizado não-supervisionado). Além disso, ela não tem a pretensão de classificar, estimar ou predizer o valor de uma variável, ela apenas identifica os grupos de dados similares;

- **Associação**: A tarefa de associação consiste em identificar quais atributos estão relacionados entre si.

2.8.1 Métodos e Técnicas: Tradicionalmente, os métodos de mineração de dados são divididos em aprendizado supervisionado (preditivo) e não

supervisionado (descritivo) (Oliveira & da Silva, 2009). Apesar do limite dessa divisão ser muito tênue (alguns métodos preditivos podem ser descritivos e vice-versa), ela ainda é interessante para fins didáticos.

A diferença entre os métodos de aprendizado supervisionados e não-supervisionados reside no fato de que os métodos não-supervisionados não precisam de uma pré-categorização para os registros, ou seja, não é necessário um atributo alvo. Tais métodos geralmente usam alguma medida de similaridade entre os atributos. As tarefas de agrupamento e associação são consideradas como não-supervisionadas.

Já no aprendizado supervisionado, os métodos são providos com um conjunto de dados que possuem uma variável alvo pré-definida e os registros são categorizados em relação a ela. As tarefas mais comuns de aprendizado supervisionado são a classificação (que também pode ser não-supervisionado) e a regressão (Oliveira & da Silva, 2009).

Durante o processo de mineração, diversas técnicas devem ser testadas e combinadas afim de que comparações possam ser feitas e então a melhor técnica (ou combinação de técnicas) seja utilizada. Assim, em plena era pós-genômica, estratégias de mineração de dados são essenciais em muitas áreas da Biologia para extrair o valor real de dados de alta vazão e, finalmente, para gerar relações úteis, regras e previsões sobre sistemas biológicos.

2.8.2 Mineração de texto: Por anos os textos tem sido a maior fonte de arquivo de informação e na atualidade a taxa na qual os artigos científicos são publicados cresce exponencialmente. De forma proporcional, cresce a necessidade de

um sistema automático que permita extrair de maneira científica a informação relevante a partir de fonte de informação primária e fundamental (Tan, 2010).

A mineração de textos é uma disciplina que junta técnicas de diversos campos como mineração de dados, lingüística, estatística computacional e ciência computacional como campos de ação. Embora a exploração de metadados é possível, a idéia básica é transformar o texto em um formato estruturado baseado em frequências de termos e assim subsequentemente aplicar técnicas conhecidas como clusterização, categorização, ontologia e análise latente de documentos por exemplo (Feinerer *et al.*, 2008)

O processo básico de uma análise de mineração de dados inclui:

- **Pré-processamento**: Que lida com a importação dos textos, a preparação, limpeza e pré-processamento em geral.

- **Associação**: Que tenta identificar associações entre termos baseadas em frequências de ocorrência e co-ocorrência.

- **Clusterização**: Que agrupa os documentos/termos em grupos de características similares.

- **Sumarização**: Que baseado na alta frequência de certos termos, os identifica como os definidores do documento.

- **Categorização**: Que classifica os documentos/textos em categorias predefinidas.

Tanto de forma comercial como na filosofia de software livre, muitas implementações para mineração de dados estão agora disponíveis, como por exemplo: Clearforest (http://www.clearforest.com/solutions.html), Summarizer

(http://www.copernic.com/en/products/summarizer/), Clementine (http://spss-clementine.software.informer.com/) entre as de uso comercial e Weka (http://www.cs.waikato.ac.nz/ml/weka/), GATE (http://gate.ac.uk/) e tm (http://cran.r-project.org/web/packages/tm/index.html) de *open source*.

2.8.3 Clusterização de variáveis: Como foi dito acima, a clusterização busca primordialmente realizar a alocação de observações, as quais são descritas por variáveis, em grupos, de forma que a similaridade seja grande entre as observações dentro de um mesmo *cluster*. Cada grupo de observações deve, assim, apresentar grande semelhança interna, ao mesmo tempo em que, se a separação dessas for adequada, as observações de um *cluster* devem ser bastante diferentes das inseridas em outro (Oliveira & da Silva, 2009).

De maneira oposta, a clusterização de variáveis visa alocar, em grupos homogêneos, precisamente as variáveis que descrevem o conjunto de observações.

Independentemente do interesse na clusterização, existem dois tipos de algoritmos para levar a cabo a análise: os algoritmos hierárquicos e os algoritmos não hierárquicos. Os algoritmos hierárquicos baseiam-se na construção de uma hierarquia entre os indivíduos, sendo esta representada graficamente através de uma estrutura conhecida como dendrograma. Os *clusters* formados são o resultado de cortes realizados nos ramos deste dendrograma (Husson et al., 2010).

Os algoritmos não hierárquicos não envolvem a construção de dendrogramas; tais técnicas agrupam as observações em *k clusters*, sendo este um valor previamente conhecido para o algoritmo, a partir da definição de centróides, que são os elementos

centrais de cada *cluster*. Esses centróides são usualmente escolhidos de forma aleatória pelos algoritmos de clusterização (Oliveira & da Silva, 2009)

Matematicamente, as observações ou as variáveis são alocadas a um determinado *cluster* de forma a minimizar a soma global das distâncias entre os membros de um *cluster* e o centróide desse *cluster*. Existem diversas métricas para calcular essa distância, sendo a distância euclidiana a mais comum (Oliveira & da Silva, 2009)

Existem outras formas de medir a similaridade de observações a serem inseridas em grupos. A distância de Manhattan, por exemplo, consiste na soma das diferenças absolutas. Outra forma de medir a similaridade pode utilizar a correlação entre as variáveis. Ao contrário das medidas baseadas em distâncias, a correlação não considera a magnitude dos valores, mas sim os padrões desses (Chavent *et al*. 2012).

Recentemente, métodos específicos baseados em correlação foram propostos para a clusterização de variáveis: CLV (https://www-admin.nantes.inra.fr/nantes_eng/les_recherches/sensometrie_et_chimiometrie/sensometrie/classification_de_variables),e ClustOfVar (http://cran.r-project.org/web/packages/ClustOfVar/index.html); sendo precisamente este tipo de clusterização o qual será abordado neste estudo.

2.8.4 Análise Fatorial: Análise fatorial é um nome genérico dado a uma classe de métodos estatísticos multivariados cujo propósito principal é definir a estrutura subjacente em uma matriz de dados. Em termos gerais, a análise fatorial aborda o problema de analisar a estrutura das inter-relações (correlações) entre um grande número de variáveis (por exemplo, escores de testes, itens de testes, respostas

de questionários), definindo um conjunto de dimensões latentes comuns, chamados fatores (Thompson, 2004).

Com a análise fatorial, o pesquisador pode primeiro identificar as dimensões separadas da estrutura e então determinar o grau em que cada variável é explicada por cada dimensão. Uma vez que essas dimensões e a explicação da cada variável estejam determinadas, os dois principais usos da análise fatorial, resumo e redução de dados - podem ser conseguidos. Ao resumir os dados, a análise fatorial obtém dimensões latentes que, quando interpretadas e compreendidas, descrevem os dados em um número muito menor de conceitos do que as variáveis individuais originais. A redução de dados pode ser conseguida calculando escores para cada dimensão latente e substituindo as variáveis originais pelos mesmos (Escofier & Pagès, 1990).

A análise fatorial desempenha um papel único na aplicação de outras técnicas multivariadas. A principal vantagem das técnicas multivariadas é sua habilidade em acomodar múltiplas variáveis em uma tentativa de compreender as relações complexas não possíveis com métodos univariados e bivariados.

Em qualquer caso, o pesquisador deve saber como as variáveis estão inter-relacionadas para melhor interpretar os resultados. Finalmente, se o número de variáveis é muito grande ou se há uma necessidade de representar melhor um número menor de conceitos, em vez das muitas facetas, a análise fatorial pode auxiliar na seleção de um subconjunto representativo de variáveis ou mesmo na criação de novas variáveis como substitutas das variáveis originais, ainda mantendo seu caráter original.

A análise fatorial difere das técnicas de dependência, nas quais uma ou mais variáveis são explicitamente consideradas como as variáveis de critério ou dependentes e todas as outras são as variáveis preditoras ou independentes.

A análise fatorial é uma técnica de interdependência nas quais todas as variáveis são simultaneamente consideradas, cada uma relacionada com todas as outras, empregando ainda o conceito da variável estatística, a composição linear de variáveis. Na análise fatorial, as variáveis estatísticas (fatores) são formadas para maximizar seu poder de explicação do conjunto inteiro de variáveis, e não para prever uma variável(eis) dependente(s). Se tiver que esboçar uma analogia com as técnicas de dependência, seria no sentido de que cada variável observada (original) é uma variável dependente que é uma função de algum conjunto latente de fatores (dimensões) feitos eles próprios a partir de todas as outras variáveis (Pàges, 2004)

Logo, cada variável é prevista por todas as outras. De maneira recíproca, pode-se olhar para cada fator (variável estatística) como uma variável dependente que é uma função do conjunto inteiro de variáveis observadas.

3. Objetivos

3.1 Objetivo geral

Este estudo tem o objetivo de identificar e quantificar a influência de caracteres genômicos nas taxas evolutivas das proteínas e descrever as possíveis associações que possam existir entre estes caracteres dentro de um sistema biológico.

3.2 Objetivos específicos

- Usar técnicas de mineração de texto para sumarizar os textos artigos na literatura tendo como base a frequência e a associação de termos.

- Coletar informação genômica sobre os termos identificados na literatura para encontrar possíveis determinantes evolutivos ou, em termos analíticos, variáveis.

- Aplicar métodos de clusterização para agrupar as variáveis de acordo com a similaridade entre elas.

- Classificar as variáveis de acordo com a natureza de cada uma delas para poder descrever o sistema evolutivo por meio de conceitos latentes que melhor descrevam a influência destas variáveis na evolução das proteínas.

- Construir um modelo de eficiência traducional que considere não só as características evolutivas de uma proteína mas também outras que, de forma global, caracterizam um sistema biológico.

4. Métodos

Este trabalho foca-se exclusivamente em genes codificados no genoma de *Saccharomyces cerevisiae*, um organismo modelo intensamente estudado e que tem uma grande disponibilidade de dados funcionais, estruturais e de expressão, constituindo assim, uma fonte de informação valiosa.

Conforme detalhado na continuação, a metodologia esta dividida em três fases principais. Uma visão mais gráfica da mesma pode ser encontrada no fluxograma incluído no ANEXO 1.

4.1 Mineração de texto

Uma busca por citações de artigos de periódicos relacionadas com o *tag* "constraints of evolution" a partir do ano 2000 foi realizada sobre o maior banco de dados de literatura científica em saúde PubMed. Títulos o resumos de citações que indicavam o interesse do estudo na identificação de fatores genômicos determinantes da evolução molecular das proteínas foram escolhidos para posterior análise. Ao final, sessenta artigos em formato PDF foram manualmente baixados do PubMed (ANEXO 2) e convertidos para arquivo de texto usando a função "pdftotext" em linux. Um código "*in-house*" implementado em linguagem C foi utilizado para processar estes arquivos extraindo as seções de interesse, tais como resumo, introdução, resultados e discussão. Os arquivos de texto resultantes formaram a coleção de documentos que

foram analisados pelo pacote "tm" (Feinerer, 2008) no ambiente R (http://www.r-project.org/) de acordo ao seguinte protocolo:

- Importação de documentos e criação do *corpus*: A estrutura principal para a análise de documentos através de técnicas de mineração de texto é aquela que permite integrar numa única instancia tanto a informação sobre cada um dos documentos (metadados) quanto o seu conteúdo (palavras). Esta estrutura é conhecida como *corpus* e para sua criação é necessário importar e identificar a uma coleção de documentos com um único corpus. No caso do presente estudo, a importação dos da coleção de documentos e sua identificação com um corpus foram realizadas usando as funções específicas do pacote "tm" para a importação de arquivos de texto (readPlain).

- Pré-processamento de documentos e termos no *corpus*: Documentos importados num corpus com sua estrutura linguística e formatação original podem ser muito difíceis de analisar por métodos de mineração de texto. Desta forma, é imprescindível a aplicação de técnicas de "limpeza" e reestruturação que podem incluir tanto a modificação dos documentos como dos termos que eles contem.

As técnicas de pré-processamento de documentos utilizadas neste estudo incluíram a remoção de números (função: *removeNumbers*), pontuação (função: *removePunctuation*), espaços em branco (função: *stripWhitespace*), palavras não importantes para o texto em inglês conhecidas como "*stopwords*" ("and", "like", "of", "on", etc).

Em quanto as técnica de pré-processamento específicas para termos, nós transformamos todas as palavras em minúsculas (função: *tolower*) e procuramos os

radicais de cada uma para reduzir a complexidade do texto sem perder informação (Stemming).

- Construção da matriz termos-documentos: Logo após o pré-processamento dos textos, a forma mais comum de apresentar os termos para posterior análise é uma matriz de termos-documentos. Esta matriz resulta da inclusão dos documentos individuais nas filas e os termos nas colunas. Conseguintemente, os elementos desta matriz correspondem as frequências de cada termo.

- Identificação de termos mais frequentes: Conceitualmente, um termo importante numa coleção de documentos é aquele que apresenta uma frequência elevada na matriz de termos-documentos. Dentro de um rango determinado, é possível identificar o conjunto de termos que poderiam estar representando a coleção de documentos; por tanto, fazendo uso da função *findFreqTerms* nós identificamos os termos que em nossa matriz de termos se repetiam pelo menos 600 vezes.

- Análise de associação entre termos mais frequentes: No simples análise de frequência é possível que alguns dos termos mais frequentes sejam verdadeiros identificadores do texto; porém, há outros que simplesmente poderiam repetir-se por questões inerentes a outros fatores metodológicos. Uma maneira mais trabalhada de encontrar os identificadores da coleção de textos é a de construir conceitos baseados nas associações existentes entre tais termos frequentes. Uma associação entre dois termos esta definida como a co-ocorrência destes dois dentro de um determinado rango de correlação. Analisando os termos mais frequentes encontrados na seção anterior, nós utilizamos a função *findAssocs* para construir conceitos genômicos em associações com correlação superior a 0.4.

4.2 Coleta de Dados

Informação referente a níveis de mRNA, eficiência traducional e abundância de proteína foi coletada para genes cujos dados comparativos de transcriptoma-proteoma estão disponíveis (MacKay *et al.*, 2004). Dados funcionais concernentes a dispensabilidade e número de interações foram obtidos de (http://chemogenomics.stanford.edu/supplements/01yfh/files/orfgenedata.txt) e da base de dados de interação de proteínas (http://dip.doe-mbi.ucla.edu/dip/), respectivamente. Informação relacionada com a estrutura nativa, percentagem de baixa complexidade e comprimento da proteína foi obtida a partir da base de dados Pedant (http://pedant.helmholtz-muenchen.de/genomes.jsp?category=fungal). Finalmente, uma função molecular foi designada a cada gene de acordo à ontologia gênica usando o SlimMapper da base de dados do genoma de *Saccharomyces* (SGD) (http://www.yeastgenome.org/).

Pares de genes ortólogos entre *Saccharomyces cerevisiae* e *Schizosaccharomyces pombe* foram encontrados usando uma versão "*stand-alone*" do algoritmo InParanoid (Ostlund *et al.*, 2010) e alinhados com o programa ClustalW 2.0 (Thompson *et al.*, 1994) com parâmetros pré-definidos. Taxas evolutivas, número de substituições não-sinônimas por sitio sinônimo (dN) e substituições sinônimas por sitio sinônimo (ds), entre cada par ortólogo, foram estimadas utilizando o método de Nei e Gojobori implementado em MEGA 4 (Tamura *et al.*, 2007).

4.3 Mineração de Dados

A sumarização pode ser visualizada como a compressão dos dados em um conjunto menor de padrões que retém ao máximo a representação da informação.

Foram utilizadas as seguintes técnicas de mineração de dados para descrever e classificar os mesmos:

4.3.1 Clusterização hierárquica de variáveis: Um algoritmo hierárquico ascendente foi usado para combinar variáveis qualitativas e quantitativas em *clusters* homogêneos. Um *cluster* de variáveis é definido como homogêneo quando as variáveis no *cluster* estão fortemente relacionadas a uma variável quantitativa sintética que representa o primeiro componente de um método de componentes principais misto (PCAMix). A pertença de uma variável em um *cluster* é definida pela correlação de razões para variáveis qualitativas e pelo coeficiente de determinação R^2 para as variáveis quantitativas. O pacote ClustOfVar implementado no ambiente R (Chavent *et al.*, 2012) foi utilizado para a execução do algoritmo.

4.3.2 Análise Fatorial Múltipla (AFM): A AFM procura a integração de grupos de variáveis que carregam informação relacionada. A análise é desenvolvida em duas etapas: Na primeira etapa, dependendo do tipo de variáveis agrupadas, análises de componentes principais (variáveis quantitativas) e/ou análises de correspondência múltipla (ACM) (variáveis qualitativas) são utilizadas para normalizar os grupos. Depois, na etapa final, uma análise de componentes principais global define a projeção dos grupos de variáveis e os fatores de carga das variáveis originais. As funções do pacote FactoMineR (Lê *et al.*, 2008) foram utilizadas para realizar a AFM em seis grupos de variáveis organizadas de acordo com a Tabela 1.

4.3.3 Análise Fatorial Bayesiana: Tendo um conjunto de variáveis observadas, a análise bayesiana de fatores incorpora um prior para a construção de um modelo que estime os índices de um fator latente. Métodos de Monte Carlo via Cadeias de Markov (Markov Chain Monte Carlo, MCMC) são utilizados para ajustar o modelo amostrando as cargas fatoriais a partir da distribuição posterior. A idéia é a de explicar através de um modelo relativamente parsimonioso as relações existentes entre um conjunto de variáveis observadas em termos de uma variável não observada (fator latente). O programa para ajustar o modelo está disponível no pacote MCMCpack (Martin *et al.*, 2011) para o ambiente R.

A perspectiva Bayesiana depende da escolha de um prior; neste caso, a restrição de uma ou mais variáveis a algum dos fatores em análise. A média e a precisão da distribuição *a priori* foram assumidas "não-informativas" com valor igual a zero. 1000 iterações iniciais foram descartadas como "queimadas" e retidas a cada 100 *scans*. 100.000 iterações foram necessárias para alcançar a convergência da Cadeia de Markov. A análise de convergência de Heidelberg e Welch foi utilizada para verificar se os valores amostrados provinham de uma distribuição estacionária.

5. Resultados

5.1 Variáveis derivadas dos identificadores de texto genômico

Uma tarefa essencial na análise de texto, inclusive no mais simples, é a de encontrar os termos que se repetem mais vezes numa coleção de documentos. Isto permite a condensação de todo o conteúdo de informações em um número limitado de palavras. Os termos frequentes representam os identificadores de uma coleção; portanto, encontrar associações significativas entre eles (isto é, termos que co-ocorrem) faz com que seja possível agrupar e organizar os conceitos a outro nível de informação mais valiosa.

Com o intuito de encontrar fatores genômicos que possam ser determinantes na evolução das proteínas, análises de frequência e associação de termos foram combinados sobre um conjunto de artigos científicos relacionados com o tema. Encontramos que trinta e um dos termos condensaram a informação do texto e alguns deles claramente caracterizavam determinantes genômicos (Figura 1).

```
[1] "chang" "correl" "data"
[4] "differ" "effect" "evolut"
[7] "evolutionari" "evolv" "express"
[10] "figur" "function" "gene"
[13] "genom" "interact" "level"
[16] "mutat" "network" "ortholog"
[19] "protein" "rate" "relat"
[22] "residu" "result" "select"
[25] "sequenc" "site" "speci"
[28] "structur" "studi" "use"
[31] "yeast"
```

Figura 1. Lista de termos mais frequentes na coleção de documentos. Trinta e uma palavras são as que resumem a informação contida em sessenta artigos científicos analisados por técnicas de mineração de texto.

Em termos de co-ocorrência, alguns destes termos apresentaram correlações significativas (Figura 2), que foram muito úteis na hora de atribuí-los a uma característica gênica ou protéica.

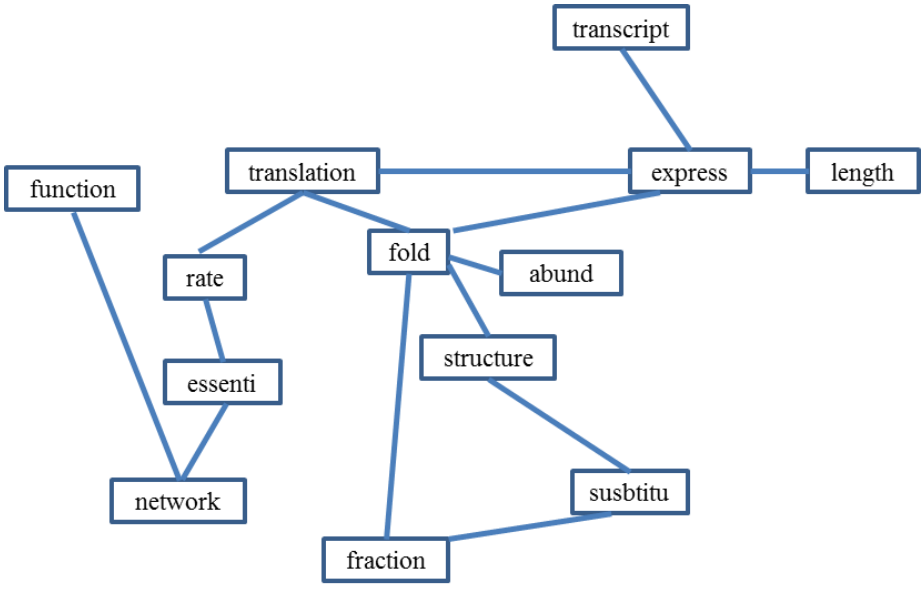


Figura 2. Rede de associação de termos. As arestas indicam a co-ocorrência significativa entre os dois vértices (termos) e providenciam a armação para o ordenamento e classificação dos termos que identificam a coleção. A identificação de associações entre termos "chave" permite a construção de conceitos biológicos relacionados com algum caráter genômico.

Ao final, treze variáveis entre características gênicas e protéicas foram identificadas e subsequentemente analisadas como potenciais determinantes da evolução de proteínas. Na Tabela 1 se apresentam os termos, o tipo de dado, a natureza e uma breve descrição das variáveis genômicas consideradas no estudo.

Tabela 1. Descrição detalhada da origem, tipo e natureza da informação genômica sintetizada a partir da análise de associação de termos frequentes.

Termo radical	Genes/Características das proteínas	Tipo de Variável	Natureza
substitu	Número de substituições sinônimas (dS)	Contínua	Evolutiva
substitu	Número de substituições não-sinônimas (dN)	Contínua	Evolutiva
express	Nível de mRNA	Contínua	Expressão
abund	Nível de proteína	Contínua	Expressão
translation	Eficiência traducional	Contínua	Expressão
length	Comprimento proteína	Contínua	Estrutural
structure	Estrutura nativa	Categórica	Estrutural
struture	Índice de instabilidade	Contínua	Estrutural
struture	Estabilidade	Categórica	Estrutural
region/structure	Percentagem de baixa complexidade	Contínua	Estrutural
network	Número de interações	Contínua	Funcional
essenti	Essencialidade	Categórica	Funcional
essenti	Dispensabilidade	Contínua	Funcional

5.2 Análises globais exploratórios revelam as relações existentes entre diferentes variáveis genômicas

Para 442 proteínas codificadas no genoma de *Saccharomyces cerevisiae* foram coletados e calculados os valores de treze variáveis genômicas construídas a partir dos termos frequentes e as associações existentes entre estes. Uma lista completa dos genes incluídos no estudo e os valores das variáveis coletadas pode ser encontrada no ANEXO 3 deste documento.

A idéia inicial foi a de analisar de forma global se estes fatores genômicos poderiam estar correlacionados com as características evolutivas ou, no caso, de expressão de cada proteína. Para isto, uma estratégia muito útil é a de construir um mapa de calor ou *heat map* que, de forma gráfica permite a visualização dos dados em forma de matriz tentando formar grupos representativos e padrões de associação em forma de tons de cores.

Como mostra a Figura 3, é possível observar a existência, em principio débil, mas estatisticamente significativa, de correlações positivas quanto positivas entre os diversos fatores incluídos no mapa de calor. É evidente por exemplo, uma correlação positiva entre a dispensabilidade de um gene com o numero de susbstituções não sinônimas acumuladas e o comprimento da proteína. Esta mesma correlação, mas negativa, é observada tanto com o nível da expressão (mRNA) quanto com a eficiência da tradução por exemplo.

Outros caracteres genômicos que foram identificados na análise de texto exibiram resultados interessantes como foi o caso do índice de instabilidade, uma variável relacionada a estrutura da proteína, que apresentou uma alta correlação positiva com dN e uma forte correlação negativa também com a eficiência traducional.

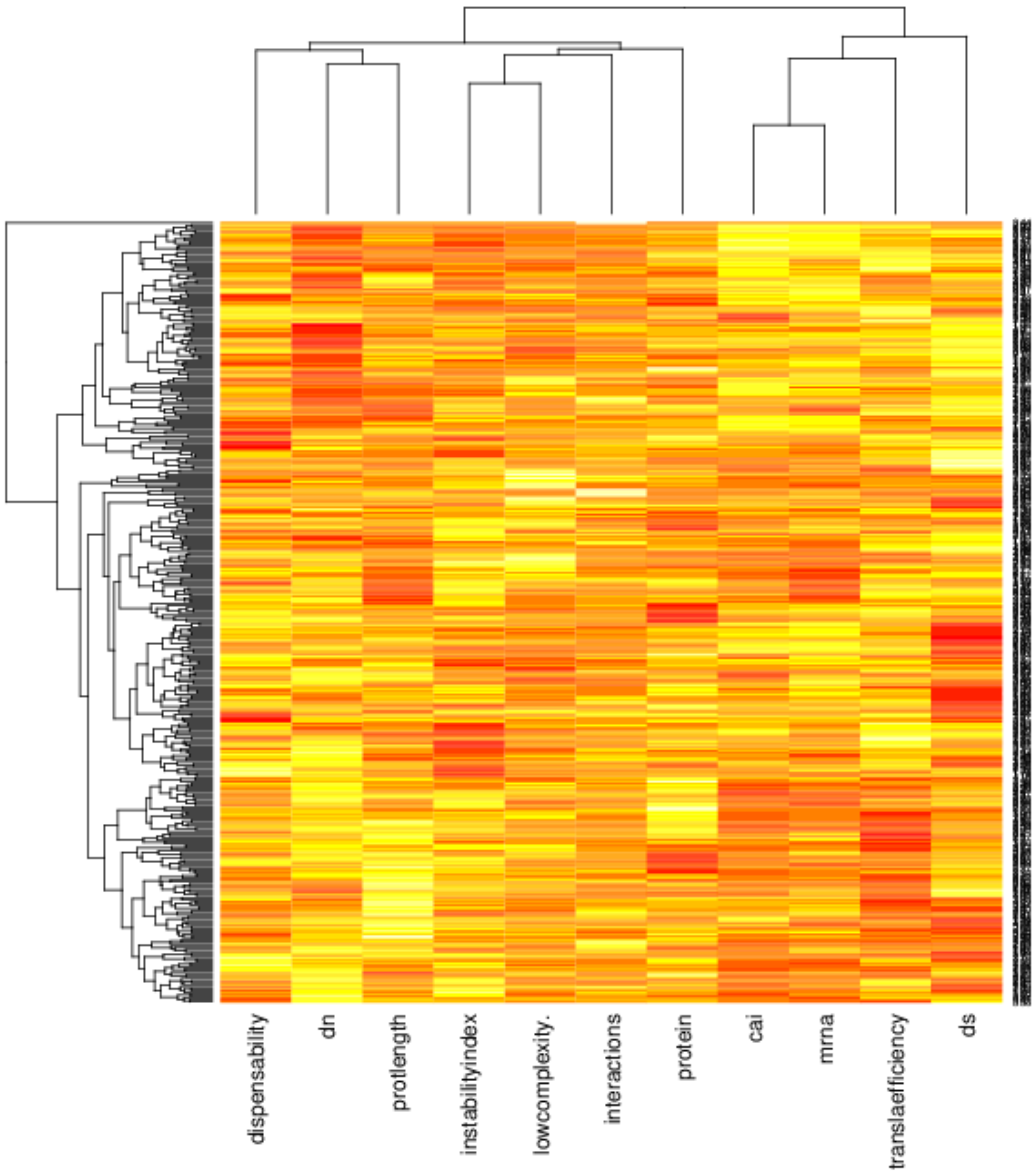


Figura 3. *Heat map* gerado a partir dos valores coletados para variáveis genômicas identificadas por análises de freqüência e associação na mineração de texto. O valor de cada variável é representado proporcionalmente ao tom de cor que permite em base ao re-ordenamento dos indivíduos no eixo da esquerda (genes) e as variáveis no eixo superior (fatores genômicos) procurar padrões de associação.

Estes resultados demonstram o potencial da mineração de texto para gerar novas informações e reforçam a noção que outros fatores genômicos que governam a evolução das proteínas existem; porém, eles ainda pouco contribuem na nossa compreensão da evolução de proteínas a partir de uma perspectiva integrada.

5.3 A clusterização de variáveis revela a estrutura dos dados

Considerando que a reunião das variáveis genômicas em grupos relacionados entre si poderia proporcionar uma perspectiva global interessante, um algoritmo de clusterização hierárquico de esquema aglomerativo foi aplicado ao grupo de variáveis composto tanto de dados quantitativos como qualitativos.

Os níveis de agregação demonstraram que quatro clusters seriam suficientes para revelar a estrutura dos dados; assim, como pode observar-se no dendrograma da Figura 4, a maioria das variáveis formou *clusters* que facilmente poderiam ser individualizados pela natureza das variáveis em cada *cluster*.

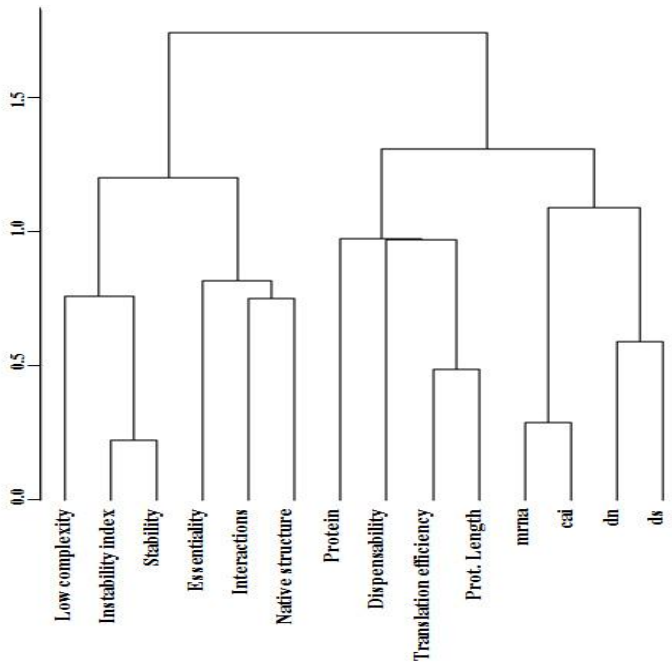


Figura 4. Clusterização hierárquica de variáveis. Duas ou mais variáveis agrupam juntas (são homogêneas) de acordo à correlação destas ao componente principal de uma variável sintética. Como resultado, quatro grupos de distintos podem ser identificados e são eles os que finalmente revelam a estrutura do conjunto de dados (C1: Percentagem de baixa complexidade, instabilidade e estabilidade estrutural. C2: Essencialidade, numero de interações e classificação estrutural. C3: Nível de proteína, dispensabilidade, eficiência de tradução e comprimento da proteína C4: Nível de RNAm, índice de adaptação do uso de códons, numero de substituições não-sinônimas e numero de substituições sinônimas).

Em termos de homogeneidade, três variáveis relacionadas com a estrutura de uma proteína, a percentagem de baixa complexidade, o índice de instabilidade e a estabilidade, claramente agruparam no mesmo *cluster*. A essencialidade e o número de interações, variáveis que poderiam ser relacionadas com a funcionalidade de uma proteína, agruparam junto com a estrutura nativa num segundo *cluster*. A abundância

da proteína, a eficiência traducional e o comprimento da proteína, intuitivamente relacionadas com a maquinaria traducional, agruparam juntos num terceiro *cluster*. Finalmente, num quarto *cluster*, as variáveis evolutivas dS e dN agruparam junto com o nível de mRNA, uma variável relacionada com a atividade gênica. As cargas fatoriais em termos de variância de cada variável no respectivo cluster podem ser encontradas na Tabela 2.

Tabela 2. Percentagens da variância explicada pelo primeiro componente principal de uma variável sintética em cada um dos clusters formados no conjunto de dados.

Cluster 1	Variância	Cluster 2	Variância
dN	0.53571258	Translation efficiency	0.72967024
dS	0.05580212	Protein level	0.06677992
mRNA	0.63455515	Dispensability	0.06813235
CAI	0.80514858	Protein length	0.70412936

Cluster 3	Variância	Cluster 4	Variância
Number of interactions	0.5174661	Low complexity	0.3843773
Essentiality	0.4435748	Instability índex	0.8422532
Native structure	0.4694705	Stability	0.7914906

5.4 Variáveis latentes são úteis para integrar dados genômicos e descrevê-los ao nível de sistemas biológicos

O agrupamento das variáveis genômicas em *clusters* permitiu compreender a estrutura subjacente do conjunto de dados; porém nenhuma informação foi fornecida sobre o tipo ou direção (positiva ou negativa) das relações existentes entre as variáveis.

Com o objetivo de analisar simultaneamente vários conjuntos de variáveis, a análise fatorial múltipla (AFM) permite reunir distintas variáveis em grupos de natureza similar para avaliar a influência de cada grupo e para revelar se existe alguma relação entre tais grupos. Logo, um conceito descritivo, conhecido como variável latente ou construto latente pode ser associado a cada um dos grupos permitindo assim atingir um novo nível de compreensão dos dados.

Seis grupos de variáveis genômicas foram criados, conforme detalhado na seção Métodos e a Tabela 1 para serem analisadas por funções incluídas no pacote FactoMineR (Lê *et al.*, 2008). A Figura 5 mostra a qualidade da representação de cada grupo de variáveis claramente separados na projeção dos eixos.

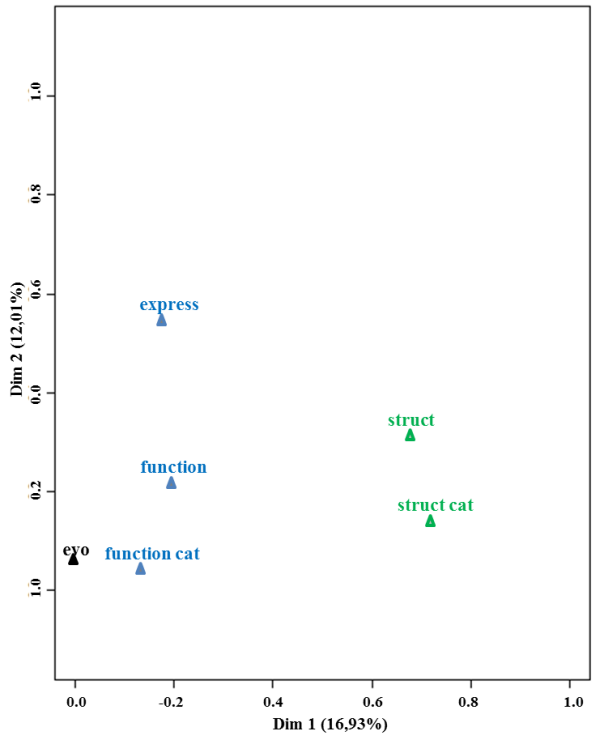


Figura 5. Representação qualitativa dos construtos latentes. Dados relacionados podem ser integrados em três principais determinantes da evolução das proteínas usando conceitos descritivos que sintetizam diferentes informações de forma confiável. Embora a acumulação da variância nos dois primeiros componentes é relativamente baixa, é possível observar que cada grupo de variáveis sintetiza um tipo de informação distinta e bem separada. No entanto variáveis relacionadas com a estrutura das proteínas tendem a se associar melhor com o primeiro componente, as variáveis relacionadas com a expressão gênica correlacionam fortemente com o segundo componente. Variáveis relacionadas com a função o rol biológico da proteína por sua parte, tendem a agrupar juntas e pouco influenciariam sobre a variabilidade dos fatores evolutivos (evo: dn e ds).

A distância entre os grupos da Figura 5 sugere que cada um deles representa informações distintas, mas integradas em três principais determinantes da evolução das proteínas: estrutura, expressão e função. Os construtos estruturais (struct e structcat) aparecem com valores fortemente coordenados com o primeiro eixo,

enquanto os construtos de expressão (express) coordenam-se claramente com o segundo eixo. Ambos os construtos estão localizados distantes do ponto de origem dos eixos e do construto "evo", que tem sido definido como grupo complementar. Isto demonstra que ambos os construtos, estruturais e de expressão, são os grupos de variáveis que mais aportam com a síntese da informação. Por outro lado, os construtos associados com função (function e functioncat), embora separados igualmente, ambos apresentaram valores baixos nos dois eixos e consequentemente apresentam pouco poder de discriminação.

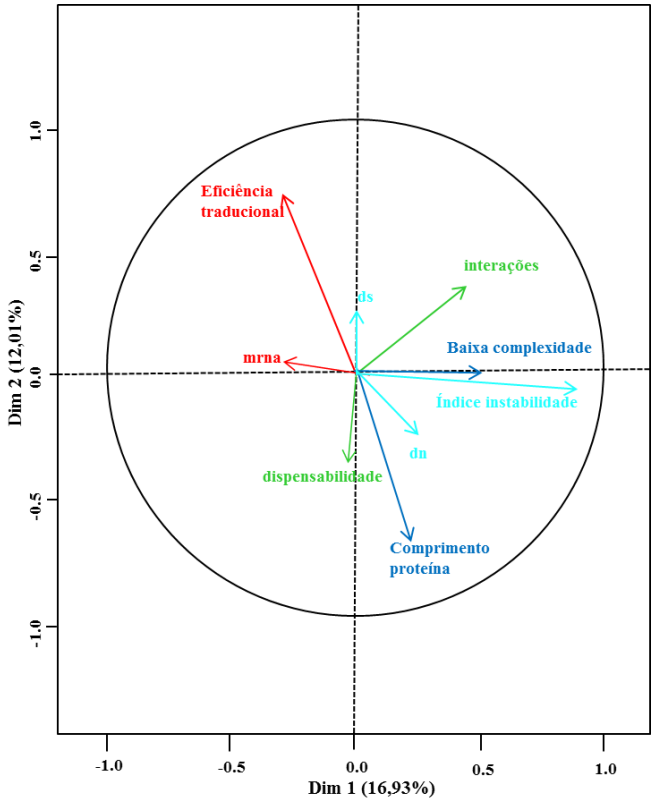


Figura 6. Círculo de correlações. Uma forma gráfica de observar as relações das variáveis ao nível individual é construir um círculo de correlações que em definitiva proporciona uma perspectiva global de um sistema. Neste círculo de correlações é possível observar as variáveis representadas por vetores cuja direção informa o tipo de associação entre duas variáveis. Variáveis correlacionadas positivamente mostram vetores na mesma direção em quanto correlações negativas são indicadas por vetores em direções opostas.

As coordenadas individuais dos membros de cada construto poderiam fornecer a perspectiva integral que descreveria um sistema biológico. Assim, a Figura 6 apresenta um mapa fatorial do circulo de correlações no qual e possível observar, por um lado, a contraposição entre as variáveis de expressão e o número de substituições não-sinônimas e, por outro lado, a alta correlação entre as variáveis relacionadas com a estrutura (percentagem de baixa complexidade e índice de instabilidade). É possível evidenciar também uma associação positiva entre a eficiência traducional e substituições sinônimas, ambas opondo-se ao comprimento de uma proteína e sua dispensabilidade.

5.5 Um modelo de fatores Bayesiano permite estimar componentes positivos e negativos de um sistema de tradução de proteínas eficiente

Para estudar os intrincados relacionamentos ao nível de um sistema em particular, uma análise de fatores Bayesianos foi utilizada sobre um conjunto de cinco variáveis genômicas: número de substituições sinônimas, eficiência traducional, abundância de proteína, dispensabilidade e índice de instabilidade, que permitiram construir os índices de um construto latente intuitivamente identificado para um sistema de tradução de proteínas. O objetivo do modelo é de capturar os padrões de associação entre as variáveis e o construto latente.

Apesar da análise Bayesiana depender de um prior, nenhuma das variáveis foi constrita para identificar o modelo e 100.000 iterações MCMC foram suficientes para alcançar uma distribuição estacionária como foi verificado pelo teste de diagnóstico (Métodos).

A Tabela 3 apresenta um resumo da distribuição posterior das cargas fatoriais e variância ou "*uniqueness*" como parte dos resultados do modelo. Em concordância

com nossas expectativas, a carga fatorial da eficiência traducional mostrou valores altos, indicando assim que existe uma forte associação entre a eficiência com que uma proteína é traduzida e o construto latente. Na mesma linha, embora mostrando uma carga fatorial relativamente menor, o número de substituições sinônimas também mostrou uma influência positiva para o que seria um sistema eficiente de tradução.

Tabela 3 Distribuição posterior das cargas fatoriais e variância de um sistema de tradução de proteínas numa análise de fatores Bayesianos. A carga fatorial mostra a correlação ou peso de cada uma das variáveis com o fator correspondente (neste caso, o primeiro e único fator).

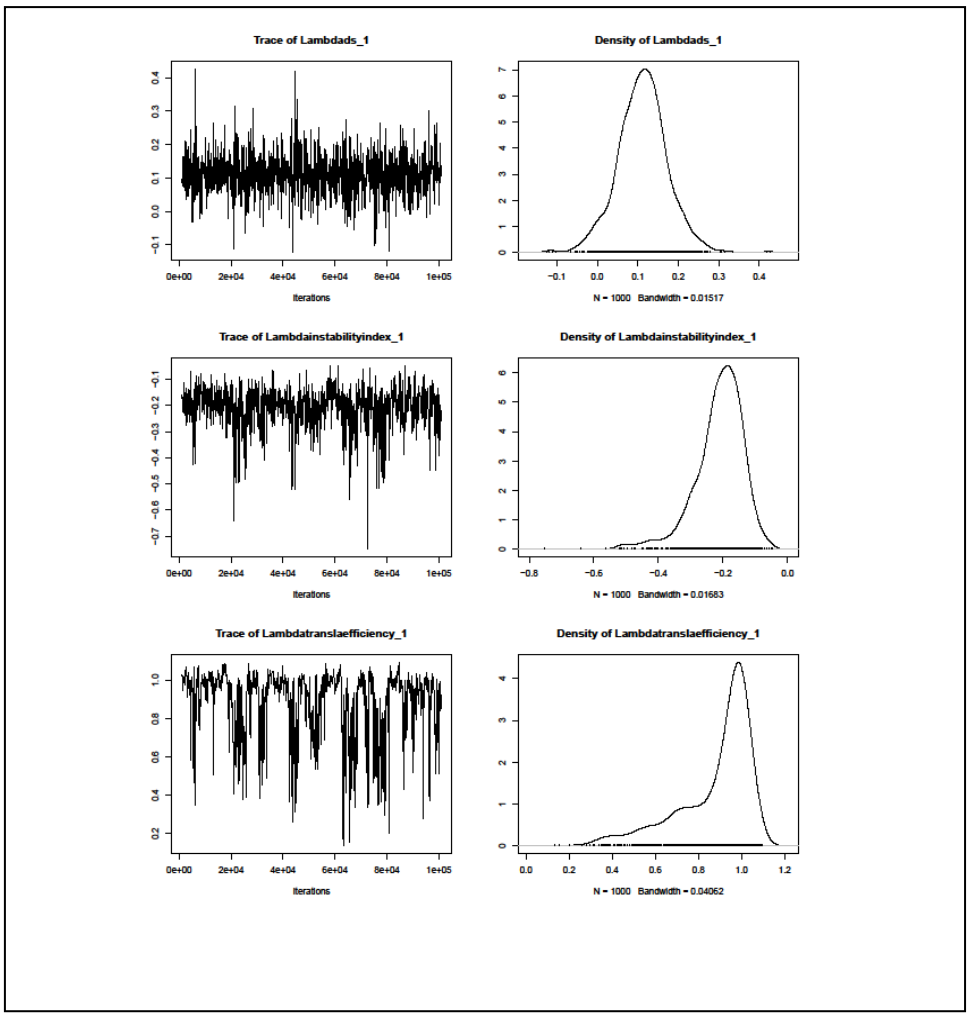
Variável	Carga Fatorial	Variância
Substituições sinônimas	0,4121	0,6921
Índice de instabilidade	-0.2134	0,9548
Eficiência tradução	0,8783	0,2129
Nível de proteína	-0.1410	0,9826
Dispensabilidade	-0.0995	0,9954

Em geral, as cargas fatoriais tendem a variar na medida em que melhor parametrizado seja o modelo, no entanto, em termos de tipo de associação, o sinal de uma carga fatorial é o que fornece a informação definitiva sobre a influência de cada variável sobre o construto latente. Desta forma e conforme a Tabela 4, o índice de instabilidade, o nível de proteína e a dispensabilidade de uma proteína foram todos estimados com cargas fatoriais negativas contribuindo assim negativamente para o sistema de tradução. Resultados do diagnóstico de convergência Tabela 4 e as densidades posteriores das variáveis no modelo são incluídos na Figura 7.

Tabela 4 Diagnóstico de convergência método Heidelberg e Welch.

Variável	Fase estacionaria	Iteração	p-value
dS	Passed	1	0.243
instability índex	Passed	1	0.180
translation efficiency	Passed	1	0.122
protein level	Passed	1	0.165
Dispensability	Passed	1	0.584

Figura 7. Distribuição das densidades posteriores das variáveis (Parcial). Os plots da Fig. 7 mostram que o numero de iterações utilizadas na análise foram suficientes para a amostragem das variáveis a partir de uma distribuição normal.



6. Discussão

Artigos científicos representam a principal fonte de informação biológica. Durante anos, os repositórios de literatura científica acumularam informação sobre atributos genômicos individuais que constituem restritores seletivos da evolução das proteínas; porém, à medida que a literatura científica cresce, maior é a necessidade por novos métodos computacionais para revelar a informação inesperada e potencialmente valiosa escondida no texto.

A mineração de texto tem surgido como uma tecnologia de avançada que, se apoiando em técnicas de recuperação de informação (RI), processamento de linguagem natural (PLN) e mineração de dados, tenta lidar com a ambiguidade da linguagem e a natureza não estruturada de documentos escritos (McDonald & Kelly, 2012). Em biologia, suas aplicações variam desde a descoberta de drogas (Plake & Schroeder, 2011), associações genéticas em doenças (Al-Mubaid & Singh, 2010) e a revisão sistemática de protocolos em biologia molecular (Krallinger *et al.*, 2005).

Como apontado anteriormente, a tarefa mais elementar numa análise de texto é extrair os termos que se repetem em uma coleção de documentos. No entanto, na prática, termos que ocorrem com baixa frequência encontram-se em poucos documentos enquanto os termos mais frequentes tendem a poluir a identificação dos principais identificadores da coleção. Portanto, o número de textos incluídos numa coleção, a transformação dos documentos, a remoção de termos contaminantes e o pré-processamento em geral constituem passos cruciais para a obtenção de resultados satisfatórios.

Inicialmente, atribuindo os identificadores da coleção de artigos com características gênicas ou protéicas, fomos capazes de revelar fatores que, à luz de

análises de correlações par-a-par, parecem constituir restrições da evolução das proteínas até agora não reconhecidos. O índice de instabilidade, a eficiência traducional e a percentagem de regiões de baixa complexidade de uma proteína estão fortemente correlacionados com a evolução acelerada de uma proteína, ou em outros termos, com o número de substituições não sinônimas (dN).

Na mesma direção, os nossos resultados mostraram que o nível de ativação de um gene, neste caso identificado pelo seu nível de mRNA, também se correlaciona negativamente com dN, apoiando a idéia de que os genes altamente expressos tendem a evoluir mais lentamente. Tem sido sugerido que a evolução progride através de alterações na expressão de proteínas (Bustamante *et al.,* 2005)e, portanto, a atividade de um gene constitui o elemento chave na nossa compreensão da evolução das proteínas.

Embora esta "chave" seja geralmente interpretada como uma associação negativa entre dN e mRNA, pode também argumentar-se que essa é uma noção excessivamente simplista do que a expressão gênica realmente representa e que restringe a seleção em uma margem de ação muito estreita. A expressão gênica pode ser explicada pelo nível ao qual um éxon é transcrito, pelo número de traduções por cada transcrito ou pelo nível de proteínas estruturalmente funcionais na célula. Deste modo, a transcrição, a tradução e a abundância de uma proteína podem ter diferentes graus de importância e a seleção natural pode ter um papel em diferentes níveis (Rocha, 2006).

A necessidade de formar e manter o sitio ativo definitivo (como ocorre no caso das enzimas) exerce uma forte pressão seletiva para uma proteína adotar apenas um dobramento estável e conservado; consequentemente, as estruturas das proteínas são

geralmente consideradas como os registros fósseis da evolução molecular (Andreeva & Murzim, 2006). No entanto, à medida que mais estruturas de proteínas tornam-se disponíveis e mais projetos de genômica estrutural geram informação nova e sem precedentes, a grande questão biológica é: Como as propriedades físicas de um sistema podem influenciar a sua capacidade de evoluir?

Por um lado, tem sido demonstrado que, contrariamente à opinião tradicional de que a função da proteína corresponde a uma estrutura tridimensional estável, muitas sequências de genes, especialmente nos genomas eucarióticos, codificam grandes segmentos ou inclusive proteínas inteiras que carecem de um enovelamento tridimensional bem definido adicionalmente, algumas destas regiões podem ser altamente conservadas entre espécies (Dyson & Wright, 2005; Nilsson & Grahn, 2011). Por outro lado, há evidências que mostram que a capacidade de algumas proteínas para evoluir é reforçada pela robustez mutacional conferida a elas graças a uma estabilidade estrutural superior (Bloom *et al.*, 2006).

Pelos exemplos precedentes podemos ver que a disponibilidade de diferentes tipos de dados biológicos serve como uma mostra da complexidade que os organismos vivos têm alcançado em milhões de anos sob a influência de forças seletivas que moldaram sua história evolutiva. O valor informativo dos dados individuais é verdadeiramente apreciado se estes estão combinados ou integrados numa única estrutura conceitual ou sistema.

As técnicas de mineração de dados podem fornecer esta estrutura e constituem uma opção ideal para a análise de conjuntos de dados diferentes, mas relacionados. Lamentavelmente, os algoritmos mais tradicionais de mineração são limitados à manipulação de dados que contêm variáveis contínuas ou categóricas, reduzindo

assim as opções do pesquisador a descartar ou discretizar uma ou outra, tornando impossível a descrição da estrutura multidimensional do conjunto de dados. Consequentemente, para explorar plenamente as características de todo o conjunto de dados, nós recorremos a métodos que são apropriados para lidar com variáveis qualitativas e quantitativas simultaneamente.

Inicialmente destinada a servir como um passo exploratório ou pré-processamento simples, a clusterização hierárquica das variáveis resultou especialmente útil para revelar a estrutura intrínseca de nossos dados. A informação que o agrupamento de variáveis traz ajuda a revelar não somente as possíveis associações entre elas, mas também facilita a compreensão de um sistema biológico, como um todo.

Os construtos latentes ou conceitos latentes desempenham um papel muito importante no trabalho teórico de muitos campos (Bollen, 2002) e aproveitamos sua virtude de atuar tanto como componentes individuais como componentes globais na explicação de um sistema, para reexaminar, à luz dos dados genômicos disponíveis, as idéias clássicas sobre a evolução das proteínas.

Uma visão clássica afirmaria que a evolução de uma proteína é basicamente governada pela seleção natural atuando sobre a estrutura e função da proteína; adicionalmente, o nível de mRNA, como "identificador" da expressão gênica, seria o maior determinante de tal evolução. Em contraposição com esta visão, a nossa abordagem prioriza a busca de determinantes globais sobre determinantes individuais.

Um processo fundamental na biologia da célula é a síntese de proteínas com elevada eficiência e fidelidade. Assim, existe um grande interesse por compreender os

mecanismos evolutivos que levaram à adaptação do sistema de tradução de proteínas (Herman *et al.*, 2012; Gilchrist *et al.*, 2009) .

O estudo de sistemas complexos, como um sistema de tradução de proteínas, começa com a identificação e a descrição simplificada dos seus componentes individuais. Uma análise fatorial Bayesiana permitiu estimar os componentes do que seria um sistema eficiente e preciso (adaptado) de tradução de proteínas. Segundo o nosso modelo, as substituições sinônimas e a eficiência traducional aportam positivamente ao sistema enquanto a dispensabilidade, o índice de instabilidade e abundância de uma proteína influenciam negativamente na adaptação do sistema.

Embora as substituições sinônimas tenham sido tradicionalmente consideradas como mostras da evolução neutra, estudos recentes demonstraram que eles exercem um efeito profundo na eficiência do sistema de tradução (Shabalina *et al.*, 2013) e também parecem influir no processo de enovelamento co-traducional das proteínas nascentes(Zhang *et al.*, 2009).

Recentemente, um estudo de Stevens *et al*. (2013) estimou a eficiência traducional para um conjunto de genes em linhas de células diferentes combinando informação referente aos níveis de mRNA e a estabilidade da proteína. De certa forma, estudos como este reforçam a linha adotada para a construção de nosso modelo.

Considerando a importância para um organismo de contar com uma suficiente disponibilidade de proteínas funcionais, a inesperada associação negativa encontrada entre a abundância da proteína e o sistema traducional eficiente, inicialmente sugere que o modelo descrito teria que ser ajustado mais apropriadamente; no entanto, esta associação negativa pode ser explicada pelo efeito de retardamento que exerce a

cinética do controle traducional através de *clusters* de códons raros que em ultima instância favorecem a fidelidade traducional sobre a eficiência traducional.

7. Conclusões

As ciências biológicas estão diante do desafio de manipular e analisar a informação biológica com a ajuda de métodos computacionais inovadores e assim responder à crescente necessidade de fazer sentido das grandes quantidades de dados experimentais. Com este fim, a integração de dados relacionados é essencial pois ela revela o verdadeiro valor do conjunto de dados e, se estiver associada a uma estrutura teórica forte, ela fornece a perspectiva global ideal para reexaminar idéias clássicas e testar novas hipóteses.

No presente capitulo, combinando técnicas de mineração de texto com simples análises de correlação, foi possível identificar características genômicas que em principio poderiam constituir determinantes da evolução das proteínas. A eficiência traduçional, a instabilidade estrutural e as regiões de baixa complexidade são tais características que puderam ser relacionadas com a taxa na qual uma proteína evolui.

Construtos latentes foram utilizados como uma alternativa para integrar dados genômicos e para abordar a evolução dos organismos biológicos como sistemas biológicos formados por componentes diferentes. O esquema de integração utilizado permitiu gerar construtos que, cada um a sua vez, claramente sintetizava uma informação específica e mostraram que, em geral, os construtos relacionados com a expressão e com a estrutura explicaram melhor o conjunto de dados em comparação com os construtos relacionados com a função. De modo geral, nossos resultados sugerem que, em vez de considerar o nível de mRNA como o maior determinante da evolução protéica, outras variáveis relacionadas com a expressão de um gene parecem ser mais importantes neste aspecto.

Um modelo de fatores Bayesiano permitiu estimar os componentes de um construto latente identificado com um sistema de tradução de proteínas eficiente. Em principio, o modelo pode carecer de rigor teórico mas, em particular, ele ajudou a compreender os padrões globais do sistema, a associação positiva entre a eficiência traducional e as substituições sinônimas e, em geral, ele demonstrou a aplicabilidade de abordagens semelhantes para a análise de outros tipos de dados biológicos.

CAPÍTULO 2

ANÁLISES DE CUSTO E BENEFÍCIO DA REGULAÇÃO CINÉTICA TRADUCIONAL

1. Introdução

O uso diferenciado de códons sinônimos, fenômeno conhecido como desvio na utilização de códons, tem sido fortemente relacionado com proteínas de alta expressão que estão envolvidas em funções celulares essenciais. Os genes que codificam estas proteínas utilizam majoritariamente códons frequentes que ainda são reconhecidos por moléculas de RNA de transferência (tRNA) em concentrações abundantes (Duret & Mouchiroud, 1999).

Desde o ponto de vista da seleção natural, a vantagem de sintetizar proteínas de forma eficiente e precisa é a força que mantém o uso diferenciado de códons sinônimos. Esta força, conhecida como seleção traducional, maximiza a velocidade de alongamento da cadeia polipeptídica, incrementa a concentração celular de ribossomos livres e minimiza a incorporação de aminoácidos errados na proteína nascente (Hershberg & Petrov, 2008; Trotta, 2013).

A seleção traducional, porém, não consegue explicar a persistência de códons não frequentes ou raros nas sequências codificantes, seu agrupamento em alguns trechos, nem o papel que eles exercem na maquinaria traducional (Komar *et al.*, 1999).

Tradicionalmente, os códons raros foram associados com um atraso na taxa de alongamento do polipeptídeo sendo sintetizado e com certas características estruturais deste, incluindo a propriedade de enovelar-se co-traducionalmente. Contudo, até há pouco, as evidências experimentais não foram suficientes nem para explicar as possíveis vantagens de manter uma proporção de códons raros nas sequências

codificantes, além do que a deriva gênica ou pressão mutacional possam explicar, nem para provar diretamente seu envolvimento no enovelamento co-traducional.

Só recentemente dois estudos, um experimental (Zhang *et al.*, 2009) e o outro fazendo uso de modelos de genética de populações (Mendez *et al.*, 2010), demonstraram que clusters de códons raros disponibilizados em alguns trechos das sequências codificantes, efetivamente tem relação com o enovelamento co-traducional de proteínas nascentes e claramente contribuem com a otimização do uso de códons na procura por uma maior aptidão.

É razoável pensar que a utilização de códons é mantida num balanço entre genes que parecem estar pressionados seletivamente para garantir um nível de proteína funcional imediato (seleção traducional) e genes que se encontram sob uma pressão seletiva exercida pela necessidade de assegurar o enovelamento co-traducional mais apropriado para a proteína (seleção cinética traducional). Portanto, a coexistência destas duas forças num genoma abre espaço a questões gerais e pontuais que ainda tem que ser exploradas. Como reconhecer a ação de uma ou de outra num organismo? Quais são os genes ou grupos de genes governados por elas?

Este capítulo apresenta uma abordagem computacional baseada numa análise de custo e benefício concebida para identificar a ação da regulação cinética traducional, as propriedades genômicas e fenômicas que poderiam definir a natureza da seleção cinética traducional e para descrever a evolução dos genes governados por ela.

2. Referencial Teórico

2.1 As proteínas como unidade funcional, estrutural e evolutiva fundamental

A maior parte do genoma dos organismos eucariotos está constituída por DNA não-codificante (90-95%) que com os anos tem demonstrado possuir importantes funções de sínteses (Non-codingfunctional RNA, por exemplo) e regulatórias (cis-elements) para a célula (Andolfatto, 2005). Porem, as proteínas (codificadas no restante 1,5-10% do genoma eucariota) ainda constituem o componente funcional e estrutural principal da maioria dos processos biológicos e resultam por tanto, elementos cruciais para o estudo da evolução dos organismos (Yang, 2009).

2.1.1 Composição química das proteínas: As proteínas são compostas por um ou mais polímeros lineares de aminoácidos ligados entre si por ligações peptídicas. Este tipo de ligação amida resulta da reação de condensação entre um grupo carboxílico alfa de um aminoácido e o grupo amino alfa de outro aminoácido. Cada cadeia pode ser chamada de peptídeo e polímeros de pequenas dimensões (tipicamente com menos de vinte aminoácidos) são denominados oligopeptídeos. Em geral, uma cadeia simples mais ou menos longa de aminoácidos é denominada polipeptídeo (Hughes, 2011).

Por serem cadeias não ramificadas, os polipeptídeos têm numa extremidade um grupo amino que não se encontra envolvido numa ligação peptídica e na outra extremidade um carboxilato nas mesmas condições. A primeira extremidade é então denominada N-terminal e a segunda C-terminal. A sequência pela qual se encontram ligados os aminoácidos é denominada **estrutura primária da proteína**, mas é mais

vulgarmente conhecida apenas por sequência de aminoácidos. Por convenção, estes são numerados começando no N-terminal, o que reflete a forma como os polipeptídeos são sintetizados na célula (também começando no N-terminal).

Como os aminoácidos perdem alguns átomos na formação da ligação peptídica, é usual denominar estes de resíduos de aminoácidos (ou simplesmente resíduos) desde o momento em que fazem parte de uma cadeia polipeptídica.

As cadeias laterais dos aminoácidos são quimicamente muito variáveis, podendo ser polares ou apolares, ionizáveis ou não, tendo diversos tamanhos e níveis de complexidade. Os milhões de possibilidades de combinação de diferentes aminoácidos que uma proteína pode ter, explica a complexidade e versatilidade das proteínas em geral (Hughes, 2011).

2.1.2 Classificação estrutural das proteínas: As proteínas possuem diferentes tipos de estrutura, além da já mencionada estrutura primária. A sequência de aminoácidos pode organizar-se espacialmente em domínios, sendo esta organização denominada **estrutura secundária**. Os principais tipos de estrutura secundária são hélices alfa e folhas beta; além destas podem referir-se os *randomcoils* (zonas desordenadas) e as beta *turn* (ligações entre folhas beta).

As hélices alfa são segmentos de polipeptídeo com uma forma em hélice em que as cadeias laterais de aminoácidos apontam para o exterior dessa hélice. Este tipo de estrutura é estabilizado pela existência de múltiplas ligações de hidrogênio no interior da hélice. Uma concentração relativamente alta de glicinas no polipeptídeo tende a forçar a existência de hélices alfa.

A estrutura em folha beta é formada por sequências do polipeptídeo que se empilham em camadas, havendo uma estabilização desta estrutura também através de

ligações de hidrogênio. As folhas podem ter uma conformação em paralelo se se encontrarem na mesma direção N-terminal—C-terminal ou em antiparalelo se empilharem em sentidos opostos. As beta *turns* ligam duas folhas beta com quatro aminoácidos numa conformação definida. Um *randomcoil* é uma zona da proteína que não tem uma estrutura secundária definida (Clark, 2012).

As proteínas adquirem a sua **estrutura terciária** ou final de forma espontânea de modo a adquirir uma configuração de energia mínima (enovelamento). *In vivo*, existem algumas proteínas (denominadas "chaperonas") que ajudam no enovelamento, especialmente quando uma proteína é muito complexa e tende a produzir conformações erradas. No entanto, a maioria das proteínas enovela-se de forma correta espontaneamente. É a estrutura primária da proteína a que determina o enovelamento final o qual pode demorar só alguns milissegundos.

Devido à enorme complexidade provocada pela existência de inúmeros aminoácidos de natureza química diversa, é difícil prever como uma proteína vai se enovelar. Porém, existem sequências de aminoácidos curtas que se repetem em diferentes proteínas e que sendo reconhecidas estruturalmente, pode-se prever como se encontrarão em outras proteínas; estas sequências são denominadas motivos.

A **estrutura quaternária** de uma proteína refere-se à presença de múltiplas cadeias polipeptídicas numa só proteína. Neste caso, diversos polipeptídeos enrolam-se formando uma proteína. O enrolamento de mais de uma cadeia numa estrutura é estabilizado pela presença de ligações químicas intermoleculares, em particular ligações dissulfeto, que ligam as diferentes cadeias numa só unidade (Clark, 2012).

2.2 A síntese de proteínas e o código genético

As proteínas não são capazes de se replicar de forma autônoma. A informação genética está contida no DNA dos cromossomos dentro do núcleo celular, mas a síntese de proteínas ocorre no citoplasma. Devido à compartimentalização das células eucarióticas, a transferência de informação do núcleo para o citoplasma é um processo muito complexo que envolve basicamente dois processos: a transcrição e a tradução. Específicamente, a tradução é o processo pelo qual o mRNA fornece um molde para a síntese de um polipeptideo; porém, o mRNA não pode se ligar diretamente a aminoácidos (Pain, 1996). É o código genético o conjunto de regras através das quais a informação contida no material genético (DNA e RNA) é traduzida em proteínas, estabelecendo-se a correspondência entre sequências de 3 nucleótidos de RNA (códons) e um determinado aminoácido.

Em teoria, são possíveis variações quase infinitas na disposição das bases ao longo de uma cadeia nucleotídica. Uma vez que existem 20 aminoácidos diferentes e apenas quatro bases diferentes de RNA, uma única base não pode especificar cada aminoácido. Em qualquer posição existem quatro possibilidades (A, T, C, G). George Gamow, utilizando o cálculo combinatório, postulou que um código de três letras (correspondente a três nucleótidos) seria necessário para codificar os 20 aminoácidos utilizados pelas células na codificação das proteínas – hipótese dos diamantes de Gamow – baseando-se no facto de existirem 4nucleótidos diferentes, combinações de 3 a 3 seriam o número mínimo para gerar mais de 20 variantes diferentes, ou seja, poderiam codificar os 20 aminoácidos existentes. A sua hipótese, embora não estivesse totalmente correta, ela serviu de base para os trabalhos posteriores

(Bollenbach, 2007). Em 1961, Nirenberg e Matthaei sintetizaram no laboratório do National Institute of Health, uma molécula de mRNA com todas as bases uracila (poli-U, isto é, uma sequencia de UUUUUUU...) e procederam à sua tradução. O polipeptídeo sintetizado consistia apenas num tipo de aminoácido, a fenilalanina. Constataram que o códon UUU era específico para o aminoácido fenilalanina. O uso de outras combinações de tripletos permitiu identificar as sequências dos códons de mRNA e os aminoácidos correspondentes, decifrando-se o código genético (Niremberg, 2004).

Dos 64 códons (RNAm) possíveis, três indicam o fim de um gene, e são conhecidos como códons finalizadores (ou sem sentido) porque designam o termino da tradução do mRNA neste ponto. São eles, o códon UAA, o UGA e o UAG. Os outros 61 especificam aminoácidos. Como existem apenas 20 aminoácidos essenciais, isto significa que a maioria dos aminoácidos pode ser especificada por mais de um códon. Por exemplo, a leucina e a arginina são especificadas por seis códons. Apenas a metionina e o triptofano são cada um deles especificado por um único códon. O código genético é, portanto, **redundante** ou degenerado (Niremberg, 2004). Embora um determinado aminoácido possa ser especificado por mais de um códon, cada códon só pode designar um aminoácido, ou seja, o código genético **não é ambíguo** (Niremberg, 2004). Essa descoberta é fundamental para, entre outras coisas, compreendermos que nem toda alteração no código genético leva a uma doença. Uma alteração de TTT para TTC, por exemplo, não deverá causar absolutamente nenhuma alteração no fenótipo de um individuo, porque ambos codificam o mesmo aminoácido. Porém há alterações na sequência de ácidos nucléicos que podem resultar em um aminoácido inapropriado sendo inserido na cadeia polipeptídica, potencialmente

causando uma doença ou mesmo a morte do organismo. Uma característica significativa do código genético é a de ser virtualmente **universal** (Niremberg, 2004), ou seja, virtualmente todos os organismos vivos usam o mesmo código para especificar aminoácidos. Uma exceção conhecida a esta regra é a das mitocôndrias, as quais têm suas próprias moléculas de DNA extranuclear. Vários códons do DNA mitocondrial codificam aminoácidos diferentes dos códons do DNA nuclear. O código genético é extremamente conservado. Os mesmos tripletes correspondem aos mesmos aminoácidos, seja em seres humanos, seja em bactérias.

2.3 O desvio de códons

O código genético é um conjunto de regras que definem a correspondência entre uma trinca de nucleotídeos (códon) no DNA e um aminoácido numa proteína. Uma característica principal do código genético é que ele é degenerado, ou seja, permite que um mesmo aminoácido seja codificado por trincas de nucleotídeos distintas, as quais são denominadas como códons sinônimos.

Já que códons sinônimos codificam para um mesmo aminoácido, é de se esperar que todos eles sejam equitativamente distribuídos ao longo das sequências codificantes num genoma, logo, sejam utilizados na mesma proporção. No entanto, códons sinônimos não estão distribuídos aleatoriamente na sequência dos genes, eles não ocorrem com a mesma frequência e consequentemente, uns são utilizados em preferência dos outros. Este fenômeno, conhecido como desvio na utilização de códons, é muito variável tanto ao nível genômico, como gênico e também intergênico (Hershberg & Petrov, 2008).

Duas visões, a princípio contrapostas, tentam explicar a origem e a evolução do desvio de códons. Por um lado, a visão selecionista sustenta que o uso preferencial de alguns códons está relacionado à eficiência e precisão na expressão das proteínas, o que supõe uma vantagem seletiva (Guoy & Gautier, 1982) e, por outro lado, a visão mutacional ou neutra, que explica a existência do desvio de códons aos padrões mutacionais de alguns dos códons que manteriam uma frequência de equilíbrio baixa (Chen *et al.*, 2014). Embora tenha sido sugerido também que um balanço entre as forças seletivas e os padrões mutacionais seria o responsável pela conservação do desvio de códons, estudos recentes mostram que a utilização preferencial de um dos códons sinônimos tem efeitos biológicos que podem refletir na aptidão do organismo (Trotta, 2013).

Neste sentido, vários fatores têm sido apontados como determinantes do uso preferencial dos códons sinônimos. O nível de expressão (Duret & Mouchiroud, 1999), a taxa de evolução (Powell & Moriyama, 1997), a estrutura secundária (Oresic & Shalloway, 1998), a localização de um gene e alguns outros podem ajudar a explicar o desvio de códons característico num determinado nível da organização genômica (Hershberg & Petrov, 2008).

Alguns índices foram desenvolvidos para quantificar o desvio de códons; entre estes, o Codon Adaptation Index (CAI) é o mais conhecido e usa um grupo de genes de referencia para determinar quais são os códons de preferência num organismo. O escore CAI para um gene é calculado a partir da frequência de todos os códons nesse gene (Sharp & Li, 1987).

2.4 A expressão gênica como determinante do desvio de códons

Expressão gênica é o processo pelo qual a informação no DNA é transcrita em RNA mensageiro (mRNA) e, depois de uma modificação pós-transcricional, traduzido pelos ribossomos para produzir uma proteína funcional. Considera-se um gene altamente expresso aquele que se ativa com frequência e que produz níveis de proteína acima da média. Por outro lado, um gene amplamente expresso é aquele que se ativa em muitas das células e tecidos de um organismo (Park & Choi, 2010).

Os genomas de uma grande variedade de organismos têm revelado uma alta correlação entre o nível de expressão gênica e o desvio de códons (Henry & Sharp, 2007; Hiraoka *et al.*, 2009). Nos genes que são traduzidos muitas vezes e em alto volume, o desvio de códons parece ser especialmente alto devido a necessidade de assegurar uma tradução eficiente e livre de erros que implicariam um elevado custo (Akashi, 1994, Akashi & Schaeffer, 1997).

Existem alguns estudos que indicam que o desvio de códons não necessariamente está restrito a genes altamente expressos (Basak *et al.*, 2008). No genoma humano, por exemplo, alguns genes de baixa expressão e outros de alta amplitude estão caracterizados por um elevado desvio de códons (Urrutia & Hust, 2001).

2.5 A seleção traducional

Como dito anteriormente, de uma perspectiva selecionista, a seleção traducional é a responsável pelo uso preferencial dos códons sinônimos.

Por um lado, uma correlação entre a frequência de um determinado códon e a abundância de seu respectivo tRNA foi demonstrada muito tempo atrás (Ikemura,

1985). Os códons mais frequentes no genoma são aqueles com maior abundância de seus respectivos tRNAs, e um marcante desvio favorecendo a utilização destes códons é encontrado em genes de alta expressão. No que se refere à seleção traducional, este desvio favoreceria a tradução eficiente de um transcrito refletindo sobre o rendimento na produção da proteína. Adicionalmente, pode gerar um benefício global à célula ao aumentar o número de ribossomos disponíveis para traduzir outras mensagens. Ao mesmo tempo, a tradução precisa e fiel do transcrito protege à célula ao reduzir o custo de metabolizar produtos errôneos, inúteis ou mesmo potencialmente tóxicos para um organismo (Hershberg & Petrov, 2008).

Tradicionalmente, a natureza da seleção traducional tem sido um tópico de grande interesse e precisamente estes dois componentes, eficiência e precisão, foram considerados em duas hipóteses: a hipótese da eficiência traducional (Qian *et al.*, 2012) e a hipótese da fidelidade traducional (Akashi, 1994; Stoletzki & Eyre-Walker, 2007) que tentam explicar as relações que existem entre o desvio de códons de um gene, seu nível de expressão e a estrutura terciária da proteína correspondente.

2.6 O enovelamento das proteínas

Peptídeos nascentes podem começar a se enovelar ainda enquanto unidos ao ribossomo num processo conhecido como enovelamento co-traducional. Durante o enovelamento co-traducional, o espaço conformacional disponível para um polipeptídeo se incrementa na medida em que mais resíduos são ligados à cadeia polipeptídica. Isto se traduz num nível adicional de controle de qualidade e um acesso a vias de enovelamento que não são possíveis para uma proteína de comprimento completo (Tourigny, 2013).

E importante notar que a cadeia linear polipeptídica é dobrada em uma estrutura tridimensional estável num período de tempo muito curto, então não é possível para a proteína sofrer muitas mudanças conformacionais até obter uma estrutura estável. Assim, foi proposto que processos controlados termodinamicamente permitem a formação de estruturas intermediárias estáveis que mais adiante irão compor a estrutura tridimensional final (Gummadi, 2003).

O conceito de paisagens de energia fornece o mecanismo pelo qual a existência de estruturas intermediárias, cada uma associada com um custo de energia livre, torna possível mapear o processo de enovelamento de uma proteína numa paisagem de energia potencial multidimensional. Ao assumir que o mapa global de energia de um enovelamento adequado apresenta a forma de funil, demonstra que só uma pequena porção de todas as estruturas possíveis consegue formar a estrutura nativa definitiva (Tourigny, 2013).

2.7 A seleção cinética traducional

Desde a sua concepção, a hipótese da seleção traducional tem sido objeto de constante questionamento pela existência de códons raros ao longo das sequências codificadoras, muito além do que a eficiência traducional poderia justificar. Estudos recentes mostraram que alguns organismos podem adaptar seu uso de códons para evitar a produção de peptídeos instáveis ou errados (Aragonès *et al.*, 2010).

Estes estudos têm sugerido que o processo de enovelamento pode ser influenciado pela cinética da tradução; assim, em contraposição à seleção traducional, dados experimentais (Zhang, *et al.*, 2009) e modelos de genética de populações (Mendez *et al.*, 2010) apóiam a existência de uma regulação cinética na tradução das

proteínas como estratégia para assegurar o enovelamento apropriado da proteína sendo sintetizada. Esta regulação é exercida através do agrupamento de códons raros (cujos tRNA´s são pouco abundantes) dispostos em trechos específicos ao longo da sequência do mRNA (Zhang *et al.*, 2009).

A utilização de códons raros para a tradução de uma proteína incrementa o tempo de emparelhamento total de seus códons com seus respectivos tRNA´s já que aqueles são pouco abundantes, o que reduz a velocidade de trânsito do ribossomo ao longo do transcrito; logo, o tempo total de síntese da proteína é maior (Komar *et al.*, 1999).

Como foi apontado anteriormente, existe uma pressão seletiva muito forte para as proteínas adotarem um enovelamento e uma estrutura tridimensional definitiva. Para isto, a exatidão e a estabilidade das estruturas intermediárias geradas durante o enovelamento co-traducional são cruciais para garantir a funcionalidade do produto final.

Ao todo, a vantagem biológica da regulação cinética exercida pela seleção cinética traducional se traduz não somente no benefício de produzir uma proteína estruturalmente estável e funcional; também se evitaria a formação de estruturas indesejadas que implicariam em um custo metabólico maior e possivelmente tóxico para a célula.

2.8 Considerações metabólicas na hipótese da eficiência traducional

De um ponto de vista energético, a síntese das proteínas é um processo muito caro (Keiron *et al.*, 2002). Por esta razão, ao longo da evolução dos genomas, as

mutações que reduzem o custo energético do processo de tradução devem ter sido favorecidas.

Os dados experimentais e as considerações precedentes demonstram que a existência dos códons raros ao longo das sequências codificadoras responde a uma necessidade de inserir pausas na tradução de uma proteína para ela testar, no espaço conformacional, as estruturas intermediárias mais estáveis.

Um atraso na cinética traducional compromete os recursos celulares que são limitados; porém, a geração veloz e imediata de peptídeos defeituosos, não funcionais e possivelmente tóxicos, também pode significar uma despesa energética e metabólica muito grande para o organismo.

Com estas considerações, é plausível um cenário, no qual existe um balanço entre o custo energético e o benefício biológico onde genes e genomas adaptam seus desvios na utilização de códons, cenário este que deve ser estudado para tentar identificar, distinguir e quantificar as forças que governam a evolução destes desvios.

3. Objetivos

Os objetivos deste trabalho podem ser resumidos nos dois itens seguintes:

- Conceber um método que permita avaliar a ação e a natureza da seleção cinética traducional.

- Identificar os genes ou grupos de genes cuja tradução possa estar submetida a uma regulação cinética, relacionada às funções biológicas que estes genes desempenham e às taxas evolutivas que os caracterizam.

4. Métodos

4.1 Taxas evolutivas

O número de substituições sinônimas por sitio sinônimo (dS) e número de substituições não sinônimas por sitio sinônimo (dN), para genes codificados no genoma de *Saccharomyces cereviseae*, no ANEXO 3 foram obtidos seguindo o protocolo descrito na seção 4.2 de Material e Métodos do capítulo precedente.

4.2 Informação estrutural e funcional

Dados relacionados com a estrutura, classificação nativa da estrutura, estabilidade, índice de estabilidade e comprimento de cada uma das proteínas foram recuperados da base de dados Mips (http://pedant.helmholtz-muenchen.de/) e SGD (http://www.yeastgenome.org/). Os genes foram classificados de acordo com a função molecular da ontologia gênica (Gene Ontology) usando o SlimMapper da SGD.

4.3 Análise custo-benefício

Taxas de alongamento individual de cada códon foram obtidas de (Gilchrist *et al.*, 2006) e códigos implementados em linguagem C foram utilizados para analisar arquivos de dados e para realizar o cálculo do custo de produção de cada proteína.

Assumindo que um gene é representado por um vetor de códons: g = {c1, c2, c3...cn} onde c*i* é o índice de alongamento do *i*th códon e n é o número de códons a ser traduzido, o cálculo da relação custo-benefício é definido por:

$$CB_{(g)} = \frac{\sum_{i=1}^{n}(c_i)}{\delta_g}$$

Onde o custo foi definido pelo tempo total de alongamento $\sum_{i=1}^{n}(c_i)$ que uma proteína demorou durante sua tradução, enquanto o benefício foi definido pelo grau de estabilidade δ_g estrutural que a proteína alcançou depois da tradução.

5. Resultados

Um balanço entre a seleção traducional e a seleção cinética traducional num genoma impõe um desafio na hora de conceber um método concreto que contextualize a vantagem de uma proteína ser definida por uma ou por outra.

Uma forma simples de abordar o problema é derivar uma relação de custo-benefício que idealmente poderia ajudar-nos a identificar os sinais de tais forças e as características dos genes ou grupos de genes governados por elas.

Uma vez definidos o custo e o beneficio associados à produção de uma proteína, o índice teria que ser avaliado, idealmente, em relação a alguma variável identificada com a estrutura da proteína. Assim, como mostra a Figura 8, uma clara diferença foi encontrada quando o custo-benefício é analisado em relação à classificação da estabilidade estrutural de uma proteína.

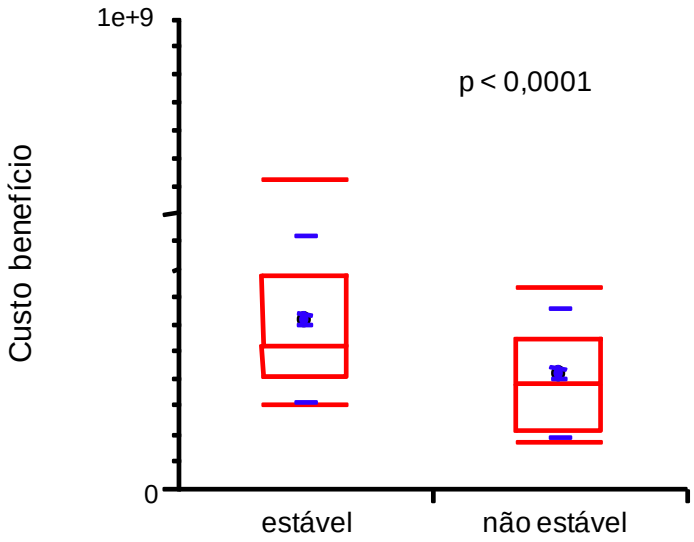


Figura 8. Box plot da relação custo-benefício e estabilidade. Diferencias apreciáveis existem entre proteínas cuja estrutura tridimensional é classificada de acordo ao grado de estabilidade. Esta diferença em relação ao custo e benefício da regulação cinética da síntese de proteínas permite identificar os grupos de genes que estariam governados por esta força.

De acordo com a seleção cinética traducional, uma maquinaria de síntese de proteína é bem adaptada se a estrutura primária, a taxa de alongamento e o processo de enovelamento co-traducional conduzem à produção de estruturas intermediárias, corretas e estáveis. Poderia se esperar então que a seleção natural promoverá a acumulação de substituições sinônimas para assegurar a adaptação de tal cinética de tradução. Na Figura 9 pode-se observar que este é aparentemente o caso. Proteínas

estáveis (presumivelmente mais adaptadas) apresentam, em média, menor número de substituições sinônimas.

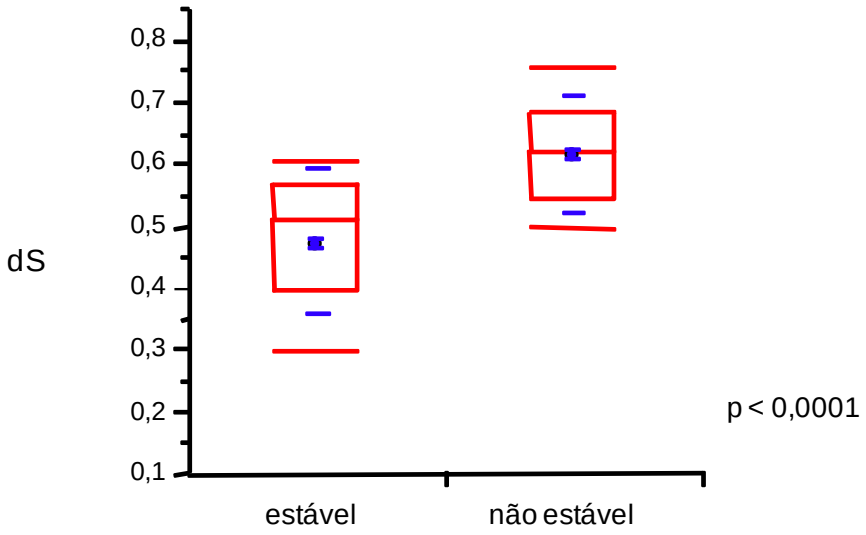


Figura 9. Box plot da acumulação de substituições sinônimas e grau de estabilidade estrutural. Os resultados sugerem que as proteínas não estáveis tendem a acumular maior numero de substituições sinônimas

Contrariamente, como mostra a Figura 10, isto não acontece no caso das substituições não-sinônimas. Não foi possível encontrar alguma diferença entre a característica estrutural de uma proteína e a acumulação de mutações que alteram os aminoácidos da mesma.

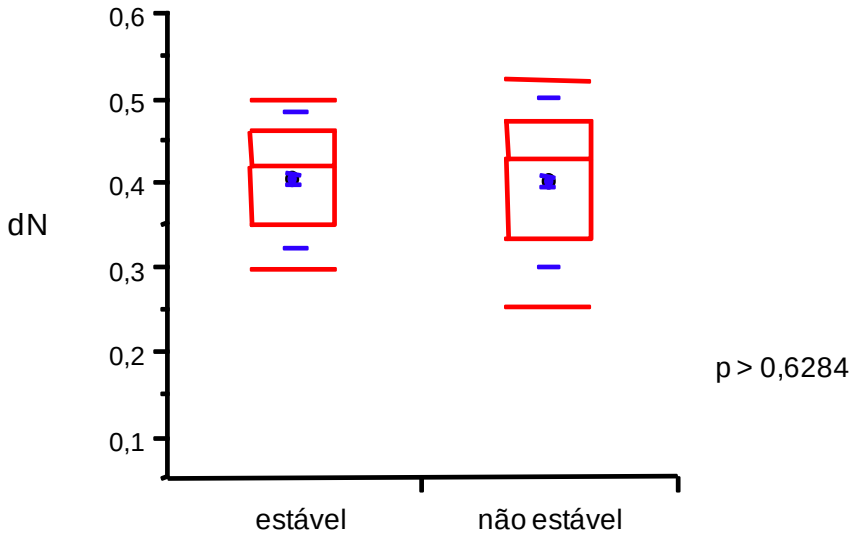
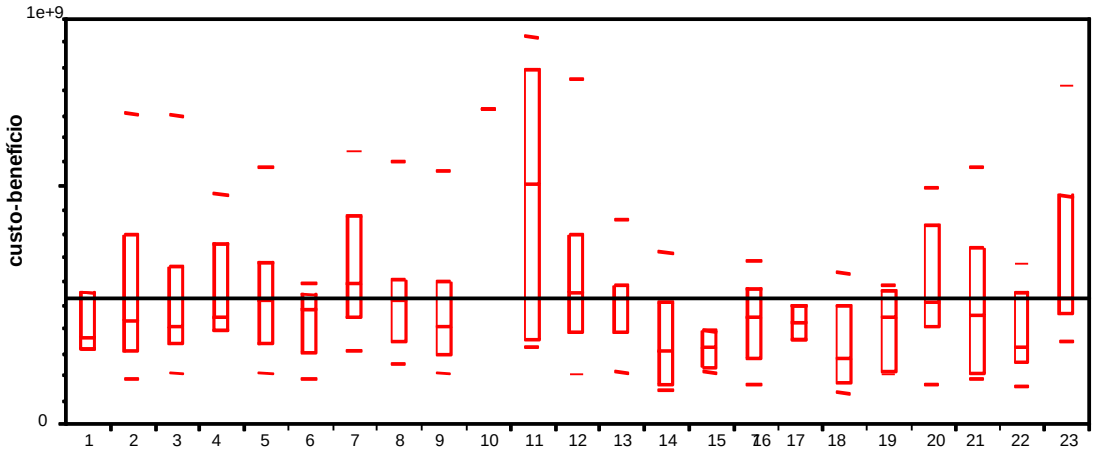


Figura 10. Box plot da acumulação de substituções não-sinônimas e o grau de estabilidade estrutural. No foi possível encontrar diferenças entre os grupos classificatórios e o numero de substituições nucleotídicas não-sinônimas.

Finalmente, com o intuito de analisar se é possível caracterizar o custo-beneficio ao nível de grupos funcionais, os genes incluídos no estudo foram identificados com uma função molecular de acordo com a classificação da ontologia gênica. A Figura 11 mostra que, embora não sejam muito pronunciadas, existem diferenças concernentes ao custo-benefício entre alguns grupos funcionais. Exemplos

destes casos são os genes com funções tais como “phosphatase phospho protein activity”, “signal transduction activity” e “transferasea ctivity”.

Figura 11. Relação custo-benefício por classificação Gene Ontology



1. DNA binding
2. RNA binding
3. Enzyme regulator activity
4. Helicase activity
5. hydrolase activity
6. isomerase activity
7. ligase activity
8. lyase activity
9. molecular function unknown
10. motor activity
11. nucleotidyl transferase activity
12. other
13. oxidoreductase activity
14. peptidase activity
15. phosphoprotein phosphatase act
16. protein binding
17. signal transducer activity
18. structural molecule activity
19. transcription regulator activity
20. transferase activity
21. translation regulator activity
22. transporter activity
23. various

6. Discussão

- Qual é a vantagem de acelerar a fase de alongamento da tradução durante a biossíntese de uma proteína?

Não e possível assegurar que uma aceleração na fase de alongamento incidirá de modo decisivo no tempo total requerido para traduzir uma proteína; isto especialmente se nós consideramos, por exemplo, que a fase de inicial da tradução poderia constituir-se num fator determinante (Hershberg & Petrov, 2008). Porém, a disponibilização rápida dos ribossomos empregados num transcrito pode incidir no nível da expressão final de um gene permitindo a reutilização destes ribossomos num outro processo e assim representar uma vantagem para a célula como um todo (Novoa & Pouplana, 2012).

- A vantagem de atrasar a fase de alongamento da tradução

Embora o uso de códons raros possa ser explicado pelos exemplos de regulação traducional através de um fenômeno conhecido como "*ribossome stalling*", dados experimentais e modelos de genética de populações tem demonstrado recentemente que um atraso na fase de alongamento da tradução é necessário para assegurar o enovelamento correto e estável de uma proteína e evitar assim a produção de estruturas não funcionais, possivelmente tóxicas e cuja proteólise provocaria um gasto metabólico adicional para a célula (Geiler-Samerotte *et al.*, 2012).

O enovelamento co-traducional é definido por uma serie de processos controlados termodinamicamente que permitem a formação de estruturas intermediárias estáveis que mais adiante irão compor a estrutura tridimensional final. Consequentemente, o requerimento de uma taxa de tradução lenta associada a uma otimização do uso dos códons também pode supor um custo energético e de recursos metabólicos que uma célula teria que disponibilizar para ter o benefício de contar com uma proteína funcional.

A análise de custo-benefício consiste num procedimento muito simples na hora de abordar problemas de tomada de decisão, principalmente nas áreas econômicas (Trebilcock *et al.,* 2007). O cenário proposto neste capitulo (a coexistência de duas forças seletivas como a seleção traducional e a seleção cinética traducional) representa para um organismo, um problema típico de tomada de decisão.

Uma etapa importante na análise de custo-benefício é definir tanto o custo como o benefício que efetivamente contextualizem o problema e, neste cenário, o beneficio deve estar identificado com alguma variável relacionada com a estrutura da proteína. O grau de estabilidade de uma proteína, identificado aqui com o benefício de atrasar a fase de alongamento da tradução, permitiu identificar grupos de proteínas caracterizados com um custo específico, embora ainda não possamos concluir que outras características vantajosas não serviriam melhor para este propósito.

Nossos resultados apóiam a existência de tal controle cinético e sugerem que proteínas instáveis, cujas relações custo-benefício são mais baixas, tendem a acumular mais substituições sinônimas, permitindo assim a exploração do espaço genotípico que providencie uma combinação de códons mais vantajosa.

Por um lado, tem sido observado que a expressão heteróloga de proteínas pode ser afetada por mudanças na cinética da tradução (Angov *et al*,. 2008); por outro lado, alguns estudos sugerem que doenças como Alzheimer, a encefalopatia espongiforme transmissível, a anemia hemolítica e outras, surgem devido a desordens de tipo conformacional das proteínas (Chaudhuri & Paul, 2006). Consequentemente, a identificação dos genes ou grupos de genes que poderiam estar governados pela regulação cinética traducional e a função biológica que eles desempenham têm implicações não somente para o campo da biotecnologia mas também para a clínica.

7. Conclusões

O estudo dos fatores envolvidos na escolha de códons e o processo de enovelamento das proteínas são temas clássicos na biologia. Relacionar ambos sob uma única hipótese é uma tarefa desafiadora.

O cenário proposto neste trabalho propõe a co-existência de duas forças: a seleção traducional e a seleção cinética traducional, cada uma das quais, a seu turno, explica o uso preferencial de um códon ou outro sinônimo de acordo com a vantagem seletiva que a identifica, eficiência ou exatidão respectivamente.

Uma abordagem de custo e beneficio foi empregada para identificar a ação da regulação cinética traducional, as propriedades genômicas e/ou características físicas que poderiam definir a natureza da seleção cinética traducional. Assim, nossos resultados mostraram diferenças significativas entre proteínas estáveis e instáveis que apoiariam a aplicação desta análise para identificar a ação da regulação cinética traducional sobre determinado grupo de genes.

As taxas evolutivas das proteínas instáveis mostraram acumular um maior número de substituições sinônimas, possivelmente uma procura no espaço genotípico pela combinação de códons mais ótima, permitindo assim reconhecer, por um lado, as marcas da pressão seletiva por manter a estrutura de uma proteína, mas ao mesmo tempo uma pressão por otimizar a cinética da sua tradução, e por outro, a natureza da seleção cinética traducional.

REFERÊNCIAS BIBLIOGRÁFICAS

- Abdi H, William LJ, Valentin D. Multiplefactoranalysis: principal componentanalysis for multitableandmultiblock data sets. WIREs Comput Stat 201.doi: 10.1002/wics.1246.

- Akashi H, Schaeffer SW. Natural selection and the frequency distributions of "silent" DNA polymorphism in *Drosophila*. Genetics 1997; 146:295-307.

- Akashi H. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics 1994; 136:927-35.

- Almeida J. *et al.* Data integration gets "sloppy". Nat Biotech 2006; 24 (9):1070-1071.

- Al-Mubaid H, Singh RK.A text-mining technique for extracting gene-disease associations from the biomedical literature.Int J Bioinform Res Appl 2010;

- Andolfatto P. Adaptive evolution of non-coding DNA in *Drosophila*. Nature 2005; 437:1149-1152.

- Andreeva A, Murzim AG. Evolution of protein fold in the presence of functional constraints. CurrOpinStructBiol 2006;16:399-408.

- Angov E, Hillier CJ, Kincaid RL, Lyon JA. Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. PLoS One. 2008; 3(5): e2189.

- Aragonès L, Guix S, Ribes E, Bosch A, Pintó RM. Fine-tuning translation kinetics selection as the driving force of codon usage bias in the Hepatitis A virus. PLoS Pathogens 2010; 6(3):e1000797.

- Arnold SJ. Constraints on phenotypic evolution.The American Naturalist. Supp. Behavioral Mechanisms in Evolutionary Ecology 1992; 140: S85-S107.

- Basak S, Mukherjee I, Chouhury M, Das S. Unusual codon usage bias en low expression genes of Vibrio cholerae. Bioinformation 2008; 3(5):213-217.

-Baxevanis A. The importance of Biological Databases in Biological Discovery.CurrProtocBioinform 2011; 34:1.1.1-1.1.6.

- Bensmail H, Haoudi A. Data Mining in Genomics and Proteomics. J Biomed Biotech 2005; 2:63-4.

- Bloom JD, Labthavikul ST, Otey CR, Arnold FH. Protein stability promotes evolvability. ProcNatlAcadSci U S A 2006; 103(15): 5869-74.

- Bollen KA. Latent variables in psychology and the social sciences.Annu Rev Psychol 2002; 53: 605-34.

- Bollenbach T, Vetsigian K, Kishony R. Evolution and multilevel optimization of the genetic code. Genome Res 2007; 17: 401-104.
- Brodie E, Moore A, Janzen F. Visualizing and quantifying natural selection. Trends EcolEvol 1995; 10(8): 313-18.
- Bulmer MG. The effect of selection on genetic variability. The American Naturalist 1971; 105(943): 201-11.
- Bustamante CD, et al. Natural selection on protein-coding genes in the human genome. Nature 2005; 437(7062): 1153-7.
- Carey G. 2003. Human Genetics for the Social Sciences.Ed. Sage publications. 2003; p. 200-33.
- Chaudhuri TK, Paul S. Protein-misfolding diseases and chaperone-based therapeutic approaches. FEBS J. 2006; 273(7): 1331-49.
- Chavent M, Kuentz-Simonet V, Liquet B, Saracco J. ClustOfVar. An R Package for the Clustering of Variables. J Statist Software 2012; 50(13):1-16.
- Chen SL, Lee W, Hottes AK, Shapiro L, McAdams H. Codon usage between genomes is constrained by genome-wide mutational processes. ProcNatlAcadSci USA 2004; 101:3480-85.
- Clark J. The structure of proteins.2012 Disponível em: http://www.chemguide.co.uk/organicprops/aminoacids/proteinstruct.html.
- Drummond A, Raval A, Wilke C, A. et al. A single determinant dominates the rate of yeast protein evolution. Mol.BiolEvol 2006;23(2): 327-337.
- Duret L, Mouchiroud D. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis, Drosophila* and *Arabidopsis*. ProcNatlAcadSci USA 1999; 96:4482-87.
- Dyson HJ, Wright PE.Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005; 6(3): 197-208.
- Escofier B, Pagès J. Multiple factor analysis. Computational Statistics & Data Analysis (1990); 18: 121–140.
- Fay JC. Sequence divergence, Functional constraint, and Selection in Protein Evolution. Annual Rev Gen Human Gen. 2003; 4:213–35.
- Fayyad U, Piatetsky-Shapiro G, Smith P, "From Data Mining to Knowledge Discovery: An Overview," U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds., Advances in Knowledge Discovery and Data Mining, pp. 1-35. AAAI/MIT Press, 1996.
- Feinerer I, An introduction to text mining in R. R News 2008; 8(2): 19-22.

- Feinerer I, Hornik K, Meyer D. Text mining infraestructure in R. Journal of Statistical Software 2008; 25(5): 1-54.
- Fernandez-Suarez XM, Galperin MY. The 2013 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection.*Nucleic Acids Research 2012;* 41(D1): D1–D7.
- Follmer C, Bezerra-Neto HJC. Fármacos multifuncionais: Monoamina oxidase e α-Sinucleína como alvos terapêuticos na doença de Parkinson. Quim Nova 2013; 36(2):306-13.
- Futuyma D. Evolution.Second Ed. Sinauer Associates. 2009; p. 279-301.
- Garcia-Diaz M, Kunkel T. Mechanism of a genetic glissando: structural biology of indel mutations. Trends BiochSci 2006; 31(4):206-14.
- Garduño R, Rein R, Egan JT, Coeckelenbergh Y, MacElroy RD. Purine-Purine Base Pairs and the Origin of Transversion-Type Mutation. Int J Quantum Chem. 1977; 4:197-204.
- Geiler-Samerotte KA, Dion MF, Budnik BA, Wang SM, Hartl DL, Drummond DA. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. ProcNatlAcadSci U S A. 2011; 108(2): 680-5.
- Gilchrist MA, Shah P, Zaretzki R. Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. Genetics 2009; 183(4): 1493-505
- Gilchrist MA, Wagner A. A model of protein translation including codon bias, nonsense errors, and ribosome recycling. J Theoretical Biol 2006; 239:417-34.
- Gopalacharyulu P. *et al*. Data integration and visualization system for enabling conceptual biology.Bionformatics 2005; 21 (1): i177-185.
- Gouy M, Gautier C. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 1982; 10:7055-74.
- Graur D, Li WH. Fundamentals of Molecular Evolution. 2nd Ed. Sinauer Associates INC, Publishers. Sunderland Massachusetts. 1999; 482 pp.
- Gummadi SN. What is the role of Thermodynamics on protein stability? Biotechnol Bioprocess Engineering 2003; 8:9-18.
- Henry I, Sharp PM. Predicting gene expression level from codon usage bias. MolBiolEvol 2007; 24(1):10-2.
- Herman D, Thomas CM, Stekel DJ. Adaptation for protein synthesis efficiency in a naturally occurring self-regulating operon.PLoS One 2012; 7(11): e49678.
- Hershberg R, Petrov D. Selection on codon bias. Annu Rev Genet 2008; 42:287-99.
- Hiraoka Y, Kawamata K, Haraguchi T, Chikashige Y. Codon usage bias is correlated with gene expression levels in the fission yeast *Schizosaccharomycespombe*. Genes to Cells 2009; 14: 499-509.

- Hughes AB. Amino acids, peptides and proteins in organic chemistry.
- Husson F, Josse J, Pages J. Principal Component Methods – Hierarchical Clustering – Partitional Clustering – Why would we need to choose for visualizing data. In <http://www.agrocampus-ouest.fr/math/>.
- Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. MolBiolEvol 1985; 2:13-34.
- Keiron P, Fraser P, Clarke A, Peck L. Low-temperature protein metabolism: seasonal changes in protein synthesis and RNA dynamics in the Antarctic limpet NacellaconcinnaStrebel 1908. J ExpBiol 2002; 205:3077-86.
- Komar AA, Lesnik T, Reiss C. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS Lett 1999; 462:387-91.
- Koonin E, Wolf Y. Evolutionary systems biology: Links between gene evolution and function. CurrOpinBiotechnol 2006;17: 481-487.
- Koonin E. Systemic determinants of gene evolution and function.Mol Sys Biol 2005; 1:2005.0021.
- Koonin EV, Wolf YI. Constraints and plasticity in genome and molecular-phenome. Nat Rev Gen 2010; 11(7): 487–98.
- Korona R. Gene Dispensability. Current Opinion Biotech 2011; 22:547-51.
- Krallinger M, Erhardt RA, Valencia A. Text-mining approaches in molecular biology and biomedicine. DDT 2005; 10(6): 439-445.
- Lacroix Z. Biological data integration: wrapping data and tools. IEEE Trans InfTechnolBiomed 2002; 6 (2): 123-128.
- Lê S, Josse J, Husson F. FactoMineR: An R Package for Multivariate Analysis. J Stat Soft 2008; 25(1): 1- 18.
- Lenormand T. Gene Flow and the limits to Natural Selection. Trends EcolEvol 2002; 17(4):183-9.
- Lercher M, Hurst L. Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet 2002; 18(7): 337-40.
- Mackay VL, *et al.* Gene expression analyzed by high-resolution state array analysis and quantitative proteomics: response of yeast to mating pheromone. Mol Cell Proteomics 2004; 3(5):478-89.
- Martin D, Quinn M, Park JH. MCMCpack: Markov Chain Monte Carlo in R. J Stat Soft 2011; 42(9): 1-21.
- McDonald D, Kelly U. The value and benefits of text mining to UK further and higher education.Digital Infraestructure JISC (2012). Disponívelem: http://bit.ly/jisc-textm.

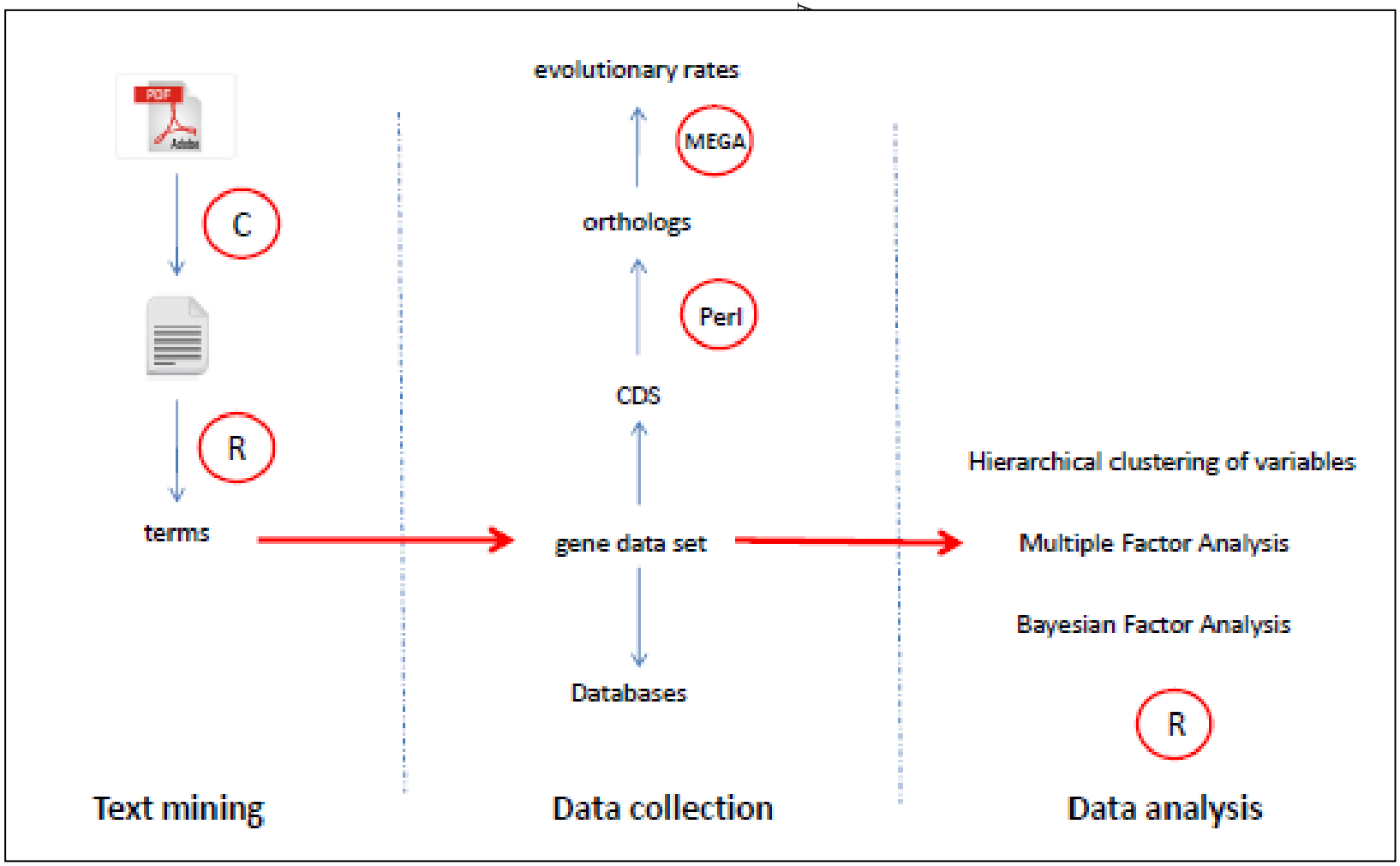
- Medina M. Genomes, phylogeny and evolutionary systems biology. PNAS 2005; 102: 6630-6635.
- Mendez R, Fritsche M, Porto M, Bastolla U. Mutation bias favors protein folding stability in the evolution of small populations. PLoS Comp Biol 2010; 6(5):e1000767.
- Mullaney J, Mills R, Pittard S, Devine S. Small insertions and deletions (INDELs) in human genomes. Hum Mol Genet 2010; 19(2): R131-R136.
- Nachman M, Crowell S. Estimate of the mutation rate per nucleotide in humans. Genetics 2000; 156(1): 297-304.
- Nachman M. Haldane and the first estimates of the human mutation rate. J Genet2004; 83(3): 231-233.
- Nilsson J, Grahn M, Wright AP. Proteome-wide evidence for enhanced positive Darwinian selection within intrinsically disordered regions in proteins. GenomeBiol 2011; 12(7):R65.
- Niremberg M. Historical review: Deciphering the genetic code—a personal account. Trends BiochemSci 2004; 29(1): 46-54.
- Novoa EM, Pouplana LR. Speeding with control: codon usage, tRNAs and ribosomes. Trends in Genetics 2012; 28(11):574-81.
- Oliveira CC, da Silva JC. Mineração de dados: Conceitos, Tarefas, Métodos e Ferramentas. Technical Report. Instituto de informática, Universidade Federal de Goiás 2009. RT-INF_001-09.
- Oresic M, Shalloway D. Specific correlations between relative synonymous codon usage and protein secondary structure. J MolBiol 1998; 281:31-48.
- Ostlund G, InParanoid 7: new algorithms and tools for eukaryotic orthology analysis, Nucleic Acids Res 2010;38: D196-203.
- Pagès J. Multiple factor analysis: Main features and application to sensory data. Revista Colombiana de Estadística 2004; 27(1): 1–26.
- Pain VM. Initiation of protein synthesis in eukaryotic cells.Eur J Biochem 1996; 236:747-71.
- Pál C, Papp B, Lercher MJ. An integrated view of protein evolution.(2006) Nat Rev Gen (2006); 7: 337-348.
- Park SG, Choi SS. Expression breadth and expression abundance behave differently in correlations with evolutionary rates. BMC Evolutionary Biology 2010; 10:241.
- Parker R. Program Abstracts Algorithms. Behavior Research methods and Instrumentation 1979; 11(3):393.
- Plake C, Schroeder M.Computational polypharmacology with text mining and ontologies. Curr Pharm Biotechnol 2011; 12(3): 449-57.

- Powell J, Moriyama E. Evolution of codon usage bias in *Drosophila*. ProcNatlAcadSci USA 1997; 94: 7784-90.
- Pray L. DNA Replication and causes of mutation. Nature Education 2008; 1(1): 214.
- Qian W, Yang JR, Pearson N, Maclean C, Zhang J. Balanced codon usage optimizes eukaryotic translational efficiency. PLoS Genet 2012; 8(3): e1002603.
- Quinn KM. Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses. Pol Anal 2004; 12:338–53.
- Rocha EP. The quest for the universals of protein evolution. Trends Genet 2006; 22(8): 412-6.
- Shabalina SA, Spiridonov NA, Kashina A Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic Acids Res 2013; 41(4) 2073–94.
- Sharp PM, Li W. The codon adaptation index-a measure of directional synonymous codon bias and its potential applications.Nucleic Acid Res 1987; 15:1281-95.
- Sherman F, Roman H. Evidence for two types of Allelic Recombination in yeast. Genetics 1963; 48(2): 255-61.
- Sniegowsky PD, Lenski RE. Mutation and Adaptation: The Directed Mutation Controversy in Evolutionary Perspective. Annu Rev EcolSyst 1995; 26:533-78.
- Stevens SG, Brown CM. In Silico Estimation of Translation Efficiency in Human Cell Lines: Potential Evidence for Widespread Translational Control. PLos One 2013; e57625. doi:10.1371/journal.pone.0057625.
- Stoletzki N, Eyre-Walker A. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. MolBiolEvol 2007; 24: 374:381.
- Subramanian S, Kumar S. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 2004; 168: 373-81.
- Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. MolBiolEvol 2007; 24(8): 1596-9.
- Tan AH. Text mining: The state of the art and the challenges. Kent Ridge Digital Labs 2010. Disponível em: http://www3.ntu.edu.sg/sce/labs/erlab/publications/papers/asahtan/tm_pakdd99.pdf.
- Thompson B. Exploratory and confirmatory factor analysis: Understanding concepts and applications. American Psychological Association.1 ed. 2004.
- Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994; 22(22): 4673-80.

- Tirosh I, Barkai N. Evolution of gene sequence and gene expression are not correlated in yeast. Trends Gen 2007; doi:10.1016/j.tig.2007.12.004.
- Torres-Reyna O. Getting Started in Factor Analysis (Using Stata 10, ver. 1.5) in <http://dss.princeton.edu/training/>.
- Tourigny D. Energy landscape theory for cotranslational protein folding. 2013.Disponívelem: arXiv:1307.6801v2.
- Trebilcock M, Yatchew A, Baziliauskas A. Overview of Cost-Benefit Analysis and its Applications in Public Policy Decisions. Market Evolution Analysis and Research Group, IESO 2007.Disponível em: https://www.ieso.ca/imoweb/pubs/mear/CRA_Overview-of-Cost-Benefit-Analysis.pdf.
- Trotta E. Selection on codon bias in yeast: a transcriptional hypothesis. Nucl Acids Res 2013; 41 (20): 9382-95.
- Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J et al. An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. Cell 2010; 141(16):344–54.
- Tzeng YH, Pan R, Li WH. Comparison of three methods for estimating rates of synonymous and nonsynonymous nucleotide substitutions.MolBiolEvol 2004;21 (12): 2290-98.
- Urrutia A, Hurst L. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in Humans, but this is not evidence for selection. Genetics 2001; 159:1191-1199.
- Wang X, Thomas SD, Zhang J. Relaxation of selective constraint and loss of function in the evolution of human bitter taste receptor genes. Hum Mol Genet 2004; 13(21): 2671-78.
- Weatherall DJ. Genotype-Phenotype relationships. Encyclopedia of Life Sciences 2001; 1:6.
- Weisbuch G. The Complex Adaptative Systems Approach to Biology. Evolution and Cognition 1999; 5(1):1-13.
- Worth CL, Gong S, Blundell TL. Structural and functional constraints in the evolution of protein families.Mol Cell Biol 2009; 10:709;20.
- Yang S, Valas R, Bourne PE. Evolution studied using protein structure. Structural Bioinformatics 2nd ed. John Wiley & Sons; 2009.
- Yao T. Bioinformatics for the genomic sciences and towards systems biology. Japanese activities in the post-genome era.ProgBiophysMolBiol 2002; 80: 23-42.
- Zhang G, Hubalewska M, Ignatova Z. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat Struct Mol Biol 2009; 16(3): 274-80.

Anexos

A.1. Fluxogramageral do estudo



A. 2 Lista de artigos analisados por técnicas de mineração de texto

Autor	Título	Periódico/Ano
Bloom, JD e col.	Structural determinants of the rate of protein evolution in yeast	Mol Biol Evol 23, 1751–61 (2006)
Brookfield, JFY	Evolution and evolvabi lity: celebrating Darwin 200	Biol Lett 5, 44–6 (2009)
Bu, L e col.	Local synteny and codon usage contribute to asymmetric sequence divergence of Saccharomyces cerevisiae gene duplicates	BMC Evol Biol 11, 279 (2011)
Chelliah, V e col.	Functional restraints on the patterns of amino acid substitutions: application to sequence-structure homology recognition	Proteins 61, 722–31 (2005)
Cowperthwaite, MC e col.	The ascent of the abundant: how mutational networks constrain evolution	PLoS Comput Biol 4, e1000110 (2008)
Drummond, DA e col.	Why highly expressed proteins evolve slowly	Proc Natl Acad Sci USA 102, 14338–43 (2005)
Drummond, DA e col.	A single determinant dominates the rate of yeast protein evolution	Mol Biol Evol 23, 327–37 (2006)
Elena, SF e col.	The effect of genetic robustness on evolvability in digital organisms	BMC Evol Biol 8, 284 (2008)
Gaucher, E e col.	Predicting functional divergence in protein evolution by site-specific rate shifts	Trends Biochem Sci 27, 315–21 (2002)
Ge, H e col.	Integrating “omic” information: a bridge between genomics and systems biology	Trends Genet 19, 551–60 (2003)
Gong, S e col.	Structural and functional restraints on the occurrence of single amino acid variations in human proteins	PLoS One 5, e9186 (2010)
Gruber, JD e col.	Contrasting properties of gene-specific regulatory, coding, and copy number mutations in Saccharomyces cerevisiae: frequency, effects, and dominance	PLoS Genet 8, e1002497 (2012)
Gu, Z e col.	Elevated evolutionary rates in the laboratory strain of Saccharomyces cerevisiae	Proc Natl Acad Sci USA 102, 1092–7 (2005)
Haerty, W e col.	Comparative analysis of function and interaction of transcription factors in nematodes: extensive conservation of orthology coupled to rapid	BMC Genomics 9, 399

	sequence evolution	(2008)
Hakes, L e col.	Specificity in protein interactions and its relationship with sequence diversity and coevolution	Proc Natl Acad Sci USA 104, 7999–8004 (2007)
Herbeck, JT e col.	Converging on a general model of protein evolution	Trends Biotechnol 23, 485–7 (2005)
Herrero, E	Evolutionary relationships between Saccharomyces cerevisiae and other fungal species as determined from genome comparisons	Rev Iberoam Micol 22, 217–22 (2005)
Hirsh, AE e col.	Protein dispensability and rate of evolution	Nature 411, 1046–9 (2001)
Hoshiyama, D e col.	Extremely reduced evolutionary rate of TATA-box binding protein in higher vertebrates and its evolutionary implications	Gene 280, 169–73 (2001)
Jordan, IK e col.	No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly	BMC Evol Biol 3, 1 (2003)
Katju, V e col.	Variation in gene duplicates with low synonymous divergence in Saccharomyces cerevisiae relative to Caenorhabditis elegans	Genome Biol 10, R75 (2009)
Kawahara, Y e col.	A genome-wide survey of changes in protein evolutionary rates across four closely related species of Saccharomyces sensu stricto group	BMC Evol Biol 7, 9 (2007)
Kim, J e col.	Rewiring of PDZ domain-ligand interaction network contributed to eukaryotic evolution	PLoS Genet 8, e1002510 (2012)
Koonin, E e col.	Evolutionary systems biology: links between gene evolution and function	Curr Opin Biotechnol 17, 481–7 (2006)
Krylov, DM e col.	Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution	Genome Res 13, 2229–35 (2003)
Larracuente, AM e col.	Evolution of protein-coding genes in Drosophila	Trends Genet 24, 114–23 (2008)
Lemos, B e col.	Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions	Mol Biol Evol 22, 1345–54 (2005)
Lin, YS e col.	Proportion of solvent-exposed amino acids in a protein and rate of protein evolution	Mol Biol Evol 24, 1005–11 (2007)
Lovell, SC e col.	An integrated view of molecular coevolution in protein-protein interactions	Mol Biol Evol 27, 2567–75 (2010)
Makino, T e col.	The evolutionary rate of a protein is influenced by features of the interacting partners	Mol Biol Evol 23, 784–9 (2006)
Makino, T e col.	Differential evolutionary rates of duplicated genes in protein interaction network	Gene 385, 57–63 (2006)
Manuscript, A	evolutionary pressures	15, 1442–1451 (2008)
McBride, RC e col.	Robustness promotes evolvability of thermotolerance in an RNA virus	BMC Evol Biol 8, 231 (2008)
McFerrin, LG e col.	The non-random clustering of non-synonymous substitutions and its relationship to evolutionary rate	BMC Genomics 12, 415 (2011)

McGuigan, K	Studying phenotypic evolution using multivariate quantitative genetics	Mol Ecol 15, 883–96 (2006)
McInerney, JO	The causes of protein evolutionary rate variation	Trends Ecol Evol 21, 230–2 (2006)
Montanari, F e col.	Differences in the number of intrinsically disordered regions between yeast duplicated proteins, and their relationship with functional divergence	PLoS One 6, e24989 (2011)
Ogurtsov, A e col.	Expression patterns of protein kinases correlate with gene architecture and evolutionary rates	PLoS One 3, e3599 (2008)
Pál, C e col.	An integrated view of protein evolution	Nat Rev Genet 7, 337–48 (2006)
Pavlicev, M e col.	Evolution of adaptive phenotypic variation patterns by direct selection for evolvability	Proc Biol Sci 278, 1903–12 (2011)
Peralta, H e col.	Sequence variability of Rhizobiales orthologs and relationship with physico-chemical characteristics of proteins	Biol Direct 6, 48 (2011)
Plotkin, JB e col.	Assessing the determinants of evolutionary rates in the presence of noise	Mol Biol Evol 24, 1113–21 (2007)
Qian, W e col.	Measuring the evolutionary rate of protein-protein interaction	Proc Natl Acad Sci USA 108, 8725–30 (2011)
Rao, YS e col.	Selection for the compactness of highly expressed genes in Gallus gallus	Biol Direct 5, 35 (2010)
Sharp, PM e col.	DNA sequence evolution: the sounds of silence	Philos Trans R Soc Lond B Biol Sci 349, 241–7 (1995)
Siegal, ML e col.	Functional and evolutionary inference in gene networks: does topology matter?	Genetica 129, 83–103 (2007)
Subramanian, S e col.	Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome	Genetics 168, 373–81 (2004)
Thorne, JL	Protein evolution constraints and model-based techniques to study them	Curr Opin Struct Biol 17, 337–41 (2007)
Tóth-Petróczy, A e col.	Slow protein evolutionary rates are dictated by surface-core association	Proc Natl Acad Sci USA 108, 11151–6 (2011)
Vieira-Silva, S e col.	Investment in rapid growth shapes the evolutionary rates of essential proteins	Proc Natl Acad Sci USA108, 20030–5 (2011)
Wall, DP e col.	Functional genomic analysis of the rates of protein evolution	Proc Natl Acad Sci USA 102, 5483–8 (2005)
Warringer, J e col.	Evolutionary constraints on yeast protein size	BMC Evol Biol 6, 61 (2006)
Wolf, Y e col.	Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution	Biol Direct 3, 40 (2008)
Wolf, Y e col.	Unifying measures of gene function and evolution	Proc Biol Sci 273, 1507–15 (2006)
Wolf, Y e col.	Relative contributions of intrinsic structural-functional constraints and translation rate to the evolution of protein-coding genes	Genome Biol Evol 2, 190–9 (2010)

Wolf, Y e col.	The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages	Proc Natl Acad Sci USA 106, 7273–80 (2009)
Yang, D e col.	An integrated view of the correlations between genomic and phenomic variables	J Genet Genomics 36, 645–51 (2009)
Yang, J e col.	Impact of translational error-induced and error-free misfolding on the rate of protein evolution	Mol Syst Biol 6, 421 (2010)
Yang, J e col.	Rate of protein evolution versus fitness effect of gene deletion	Mol Biol Evol 20, 772–4 (2003)
Zhang, J e col.	Significant impact of protein dispensability on the instantaneous rate of protein evolution	Mol Biol Evol 22, 1147–55 (2005)
Zhou, T e col.	Contact density affects protein evolutionary rate from bacteria to animals	J Mol Evol 66, 395–404 (2008)

A.3 Lista de genes e valores de variáveis incluídos no estudo

systematic name	dn	ds	dn/ds	mRNA	transla efficienc y	Protein	cai	inter actio ns	dispens ability	essentia lity	low comple xity %	prot length	instabili ty index	stability	native structure	GO
YAL003W	0.292	0.554	0.527075812	604.33	7.215	0.982318271	0.741	3	0.987	YES	23.8	206	42.13	unstable	alpha beta	translation regulator activity
YAL016W	0.394	0.629	0.626391097	67.4	4.636	0.612369871	0.177	16	0.991	NO	4.3	635	41.95	unstable	all alpha	phosphoprotein phosphatase
YAL025C	0.313	0.731	0.428180575	10.63	5.297	1.097.694.841	0.219	1	1.017	YES	30.7	306	59.8	unstable	all alpha	molecular function unknown
YAL035W	0.352	0.627	0.561403509	371.03	3.885	110.864.745	0.355	17	0.987	NO	28.5	1002	48.89	unstable	all alpha	translation regulator activity
YAL038W	0.253	0.53	0.477358491	5613.56	5.331	1.199.040.767	0.893	2	0.99	YES	3	500	23.23	stable	alpha beta	transferase activity
YAL039C	0.382	0.76	0.502631579	12.75	4.043	0.925925926	0.114	0	1.025	NO	5.2	269	61.44	unstable	alpha beta	lyase activity
YAL042W	0.468	0.656	0.713414634	84.87	6.613	0.912408759	0.118	3	0.989	NO	0	415	33.52	stable	Membrane	molecular function unknown
YAL062W	0.358	0.558	0.641577061	61.83	5.4	0.839630563	0.156	2	1.02	NO	3.9	457	24.73	stable	alpha beta	oxidoreductase activity
YBL008w	0.491	0.479	1.025.052.192	6.19	2.084	1.428.571.429	0.128	6	0.992	NO	1.1	840	42.68	unstable	all beta	protein binding
YBL017c	0.49	0.623	0.786516854	5.28	0.163	1.027.749.229	0.163	4	1.003	NO	4.3	1579	35.4	stable	Membrane	Other
YBL024w	0.379	0.699	0.542203147	122.75	4.959	1.097.694.841	0.27	4	1.007	NO	9.2	684	40.42	unstable	alpha beta	transferase activity
YBL036c	0.418	0.71	0.588732394	12.05	6.541	1.046.025.105	0.236	7	0.991	NO	4.7	257	26.77	stable	alpha beta	isomerase activity
YBL039c	0.257	0.739	0.347767253	74.97	5.013	1.100.110.011	0.309	21	1.001	NO	4.7	579	34.87	stable	alpha beta	ligase activity
YBL050w	0.4	0.667	0.59970015	17.11	5.071	0.970873786	0.159	36	1.005	YES	3.1	292	30.01	stable	all alpha	protein binding
YBL072c	0.233	0.622	0.374598071	2521.56	3.452	0.860585198	0.747	0	0.996	NO	14	200	52.34	unstable	alpha beta	structural molecule activity
YBL076c	0.347	0.46	0.754347826	605.13	3.917	127.388.535	0.342	8	1.012	YES	2.9	1072	35	stable	alpha beta	ligase activity
YBL079w	0.479	0.547	0.875685558	2156.69	0.394	0.800640512	0.151	3	1.021	NO	2.5	1502	42.12	unstable	alpha beta	structural molecule activity
YBL087c	0.156	0.666	0.234234234	482.64	4.524	0.986193294	0.624	0	0.958	NO	0	137	32.02	stable	all beta	structural molecule activity
YBL091c	0.273	0.758	0.360158311	29.37	4.698	0.854700855	0.211	1	0.913	NO	4.3	175	33.25	stable	alpha beta	peptidase activity
YBR025c	0.276	0.753	0.366533865	766.37	5.659	1.196.172.249	0.567	8	0.984	NO	0	394	35.83	stable	alpha beta	hydrolase activity
YBR031w	0.263	0.626	0.420127796	4616.87	6.488	0.637755102	0.803	0	0.982	NO	11.3	362	33.11	stable	alpha beta	structural molecule activity
YBR034c	0.396	0.42	0.942857143	176.11	7.148	0.965250965	0.267	9	1.004	NO	0	348	38.34	stable	alpha beta	transferase activity
YBR048w	0.208	0.655	0.317557252	1700.92	2.524	1.082.251.082	0.733	0	0.984	NO	0	156	53.58	unstable	all beta	structural molecule activity
YBR058c	0.457	0.576	0.793402778	15.53	3.076	1.088.139.282	0.162	10	0.986	NO	6.3	781	40.95	unstable	alpha beta	hydrolase activity
YBR078w	0.503	0.463	1.086.393.089	2898.32	5.055	0.953288847	0.553	1	0.887	NO	27.1	468	32.68	stable	Membrane	molecular function unknown
YBR082c	0.126	0.685	0.183941606	34.73	8.002	1.206.272.618	0.313	2	0.989	NO	0	148	48.24	unstable	alpha beta	ligase activity
YBR087w	0.419	0.46	0.910869565	22.93	3.816	1.754.385.965	0.152	13	1.003	YES	0	354	39.86	stable	all alpha	DNA binding
YBR101c	0.47	0.686	0.685131195	58.09	5.939	0.930232558	0.158	14	0.89	NO	0	290	46.21	unstable	all alpha	Other
YBR115c	0.407	0.444	0.916666667	24.32	3.315	1.067.235.859	0.212	2	0.797	NO	3.2	1392	28.57	stable	alpha beta	oxidoreductase activity
YBR121c	0.307	0.729	0.421124829	216.03	5.018	0.928505107	0.414	3	1.004	YES	7.8	667	34.19	stable	alpha beta	ligase activity
YBR127c	0.344	0.121	2.842.975.207	689.74	5.95	0.966183575	0.39	10	1.014	NO	3.7	517	33.72	stable	alpha beta	hydrolase activity
YBR133c	0.499	0.567	0.880070547	54.85	1.838	1.545.595.054	0.127	8	0.965	NO	1.8	827	48.38	unstable	alpha beta	transferase activity
YBR143c	0.245	0.687	0.356622999	394.48	4.85	1.038.421.599	0.334	21	0.976	YES	3.9	437	25.56	stable	alpha beta	translation regulator activity
YBR162c	0.351	0.634	0.55362776	1639.34	5.251	0.874890639	0.381	10	0.969	NO	12.7	455	45.15	unstable	all alpha	molecular function unknown
YBR234c	0.473	0.572	0.826923077	141.44	4.663	1.121.076.233	0.197	16	0.978	YES	0	384	28.72	stable	all beta	structural molecule activity
YBR237w	0.505	0.496	1.018.145.161	18.35	2.525	0.297885016	0.131	1	1	YES	10.7	849	48.18	unstable	all alpha	Various

YBR248c	0.393	0.364	107.967.033	18.05	5.431	1.251.564.456	0.16	1	1.022	NO	2.4	552	35.61	stable	alpha beta	transferase activity
YBR249c	0.4	0.34	1.176.470.588	793.95	6.003	1.082.251.082	0.527	5	0.99	NO	0	370	34.92	stable	alpha beta	transferase activity
YBR265w	0.547	0.577	0.948006932	37.43	4.671	0.731528895	0.15	3	0.984	YES	0	320	44.62	unstable	Membrane	oxidoreductase activity
YCL009c	0.394	0.694	0.567723343	547.84	2.945	1.092.896.175	0.242	4	0.98	NO	0	309	57.11	unstable	alpha beta	transferase activity
YCL011c	0.558	0.64	0.871875	61.72	3.552	0.991080278	0.168	13	0.99	NO	20.6	427	44.7	unstable	alpha beta	RNA binding
YCL017c	0.436	0.286	1.524.475.524	137.01	4.571	0.914076782	0.226	3	0.948	YES	8.5	497	34.03	stable	alpha beta	lyase activity
YCL030c	0.427	0.436	0.979357798	110.55	4.512	108.577.633	0.269	5	0.952	NO	1.6	799	36.7	stable	alpha beta	hydrolase activity
YCL043c	0.453	0.494	0.917004049	1591.53	5.454	0.484027106	0.404	7	0.962	YES	9.6	522	40.3	unstable	alpha beta	isomerase activity
YCR033w	0.514	0.594	0.865319865	38.7	0.579	0.89206066	0.12	8	0.992	NO	12.9	1226	58.63	unstable	alpha beta	hydrolase activity
YCR053w	0.333	0.587	0.567291312	362.24	5.429	1.104.972.376	0.404	4	1.014	NO	2.3	514	31.94	stable	all alpha	lyase activity
YCR084c	0.414	0.661	0.626323752	357.62	3.51	1.280.409.731	0.181	15	0.994	NO	18.5	713	45.8	unstable	alpha beta	transcription regulator activity
YDL014W	0.209	0.6	0.348333333	1085.54	5.272	0.856164384	0.492	31	0.985	YES	24.2	327	37.01	stable	all beta	transferase activity
YDL022w	0.354	0.767	0.461538462	78.3	7.299	0.996015936	0.46	1	0.999	NO	0	391	30.91	stable	alpha beta	oxidoreductase activity
YDL029W	0.221	0.787	0.280813215	96.2	5.362	0.615384615	0.209	48	0.998	YES	3.1	391	40.6	unstable	alpha beta	protein binding
YDL043C	0.543	0.6	0.905	6.28	5.534	1.278.772.379	0.153	40	0.983	YES	11.7	266	44.28	unstable	all alpha	RNA binding
YDL046w	0.477	0.687	0.694323144	172.69	4.252	1.131.221.719	0.228	2	0.993	NO	5.8	173	37.95	stable	Membrane	molecular function unknown
YDL051W	0.431	0.502	0.858565737	52.29	3.475	103.950.104	0.25	9	0.998	NO	16.4	275	63	unstable	alpha beta	RNA binding
YDL055C	0.259	0.519	0.499036609	2124.64	5.111	1.074.113.856	0.6	2	0.999	YES	0	361	25.53	stable	alpha beta	transferase activity
YDL060w	0.456	0.522	0.873563218	59.7	3.931	1.016.260.163	0.182	16	0.985	YES	4.4	788	43.95	unstable	alpha beta	Other
YDL066W	0.256	0.741	0.345479082	395.93	6.637	1.538.461.538	0.319	1	1	NO	6.8	428	33.13	stable	alpha beta	oxidoreductase activity
YDL084w	0.306	0.478	0.640167364	756.13	4.078	1.141.552.511	0.374	5	0.991	YES	4.3	446	32.8	stable	all alpha	Various
YDL095W	0.511	0.501	101.996.008	471.95	5.486	0.950570342	0.227	1	0.991	NO	0	817	42.64	unstable	Membrane	transferase activity
YDL097c	0.472	0.363	1.300.275.482	55.29	4.419	0.952380952	0.154	30	0.973	YES	7.4	434	40.75	unstable	all alpha	structural molecule activity
YDL100c	0.426	0.25	1.704	79.04	5.259	1.388.888.889	0.322	32	0.993	NO	9.6	354	30.55	stable	alpha beta	hydrolase activity
YDL102W	0.322	0.626	0.514376997	14.06	4.124	0.448028674	0.176	2	0.974	YES	1.3	1097	38.27	stable	alpha beta	Various
YDL111c	0.491	0.551	0.891107078	20.45	2.79	1.239.157.373	0.116	9	0.985	YES	0	265	34.87	stable	alpha beta	hydrolase activity
YDL116W	0.53	0.57	0.929824561	56.33	4.674	0.856898029	0.15	31	0.98	NO	0	726	42.99	unstable	all alpha	structural molecule activity
YDL124w	0.436	0.671	0.649776453	52.56	5.317	0.935453695	0.197	2	0.986	NO	0	312	48.57	unstable	alpha beta	oxidoreductase activity
YDL126C	0.214	0.592	0.361486486	437.12	4.831	1.141.552.511	0.307	13	0.99	YES	11	835	30.83	stable	alpha beta	hydrolase activity
YDL131w	0.181	0.8	0.22625	284.42	6.484	0.997008973	0.329	1	0.989	NO	1.8	440	33.94	stable	alpha beta	transferase activity
YDL134C	0.214	0.798	0.268170426	188.41	3.749	0.899280576	0.146	16	1.005	NO	4.6	369	33.56	stable	alpha beta	phosphoprotein phosphatase
YDL143W	0.37	0.369	1.002.710.027	276.09	5.066	1.089.324.619	0.225	2	0.989	YES	2.7	528	41.79	unstable	alpha beta	protein binding
YDL145C	0.42	0.518	0.810810811	329.24	3.259	1.133.786.848	0.237	30	0.983	YES	3	1201	32.11	stable	alpha beta	molecular function unknown
YDL160C	0.246	0.737	0.333785617	157.66	5.683	1.126.126.126	0.21	19	0.978	NO	8.9	506	47.06	unstable	all alpha	helicase activity
YDL166c	0.356	0.772	0.461139896	5.94	6.374	0.899280576	0.146	3	0.995	YES	11.2	197	47.23	unstable	all alpha	hydrolase activity
YDL167C	0.526	0.637	0.825745683	70.94	2.276	1.584.786.054	0.136	3	1.004	NO	19.2	719	48.5	unstable	alpha beta	molecular function unknown
YDL168W	0.239	0.749	0.319092123	51.28	5.987	0.869565217	0.243	1	1.011	NO	6.2	386	27.35	stable	alpha beta	oxidoreductase activity
YDL171c	0.332	0.503	0.660039761	351.92	2.173	1.083.423.619	0.287	5	0.983	NO	3	2145	32.38	stable	alpha beta	oxidoreductase activity
YDL185W	0.425	0.397	1.070.528.967	544.16	4.158	1.052.631.579	0.305	31	0.993	NO	1.2	1071	33.01	stable	alpha beta	hydrolase activity
YDL201w	0.3	0.676	0.443786982	20.56	5.345	0.703729768	0.19	1	0.988	NO	0	286	37	stable	alpha beta	protein binding
YDL236W	0.43	0.598	0.719063545	71.17	6.946	0.871080139	0.196	1	0.963	NO	4.8	312	34.74	stable	alpha beta	hydrolase activity
YDR002W	0.35	0.712	0.491573034	79.76	7.103	1.055.966.209	0.489	7	0.88	YES	18.9	201	44.28	unstable	all beta	protein binding
YDR005C	0.454	0.499	0.909819639	31.63	2.923	1.782.531.194	0.119	4	0.978	NO	17.5	395	54	unstable	alpha beta	transcription regulator activity
YDR011W	0.415	0.57	0.728070175	131.95	2.385	1.161.440.186	0.18	0	0.967	NO	5.3	1501	39.82	stable	Membrane	hydrolase activity
YDR023W	0.209	0.74	0.282432432	445.01	5.878	1.091.703.057	0.392	9	0.978	YES	8.4	462	41.22	unstable	alpha beta	ligase activity
YDR037W	0.414	0.23	1.8	803.64	5.517	0.845308538	0.422	6	0.985	YES	3.2	591	41.12	unstable	alpha beta	ligase activity
YDR047W	0.28	0.837	0.334528076	81.61	4.837	0.712250712	0.16	0	0.972	YES	0	362	37.56	stable	alpha beta	lyase activity

YDR050C	0.328	0.57	0.575438596	2599.11	6.605	0.972762646	0.817	1	1.014	YES	0	248	19.66	stable	alpha beta	isomerase activity
YDR060w	0.403	0.671	0.600596125	23.35	3.335	0.607164542	0.2	28	0.985	YES	13.5	1025	42.3	unstable	all alpha	molecular function unknown
YDR061w	0.496	0.564	0.879432624	13.73	3.37	0.085638435	0.099	2	1.002	NO	1.5	539	48.37	unstable	alpha beta	transporter activity
YDR071c	0.442	0.653	0.676875957	10.49	8.044	0.62305296	0.244	6	0.733	NO	0	191	47	unstable	alpha beta	transferase activity
YDR091C	0.2	0.723	0.276625173	479.04	4.765	119.760.479	0.369	8	1.001	YES	0	608	36.65	stable	alpha beta	hydrolase activity
YDR101C	0.467	0.553	0.844484629	233.34	5.508	1.031.991.744	0.238	15	0.988	NO	0	593	40.68	unstable	alpha beta	molecular function unknown
YDR120C	0.45	0.479	0.939457203	56.04	6.155	0.950570342	0.155	4	0.987	NO	1.9	570	48.17	unstable	alpha beta	transferase activity
YDR129C	0.256	0.736	0.347826087	183.5	5.812	0.881057269	0.234	8	0.977	NO	2	642	38.28	stable	all alpha	protein binding
YDR152W	0.515	0.605	0.851239669	9.14	4.932	1.023.541.453	0.182	3	1.028	NO	16.6	265	35.38	stable	alpha beta	molecular function unknown
YDR158W	0.338	0.571	0.591943958	1155.74	4.979	1	0.431	4	1.006	NO	0	365	39.08	stable	alpha beta	oxidoreductase activity
YDR161W	0.532	0.554	0.960288809	33.75	3.173	1.408.450.704	0.122	0	1.001	NO	8.3	387	44.29	unstable	all alpha	molecular function unknown
YDR170C	0.471	0.53	0.888679245	107.2	2.055	1.410.437.236	0.193	23	1.004	YES	8.8	2009	46.64	unstable	all alpha	enzyme regulator activity
YDR172W	0.431	0.302	1.427.152.318	86.04	4.819	0.92936803	0.315	13	1.023	YES	27.9	685	41.43	unstable	alpha beta	translation regulator activity
YDR188W	0.406	0.397	1.022.670.025	388.83	4.531	1.152.073.733	0.177	18	1.011	YES	0	546	37.4	stable	alpha beta	protein binding
YDR190C	0.218	0.727	0.299862448	54.17	5.47	1.074.113.856	0.19	17	0.996	YES	0	463	42.19	unstable	alpha beta	helicase activity
YDR211W	0.453	0.411	1.102.189.781	67.22	3.846	1.041.666.667	0.198	10	1.013	YES	5.5	712	47.33	unstable	alpha beta	translation regulator activity
YDR212W	0.173	0.805	0.214906832	426.17	5.404	0.877192982	0.244	22	1.01	YES	4.5	559	40.37	unstable	alpha beta	protein binding
YDR234W	0.263	0.75	0.350666667	148.11	4.157	0.8	0.2	0	0.999	NO	6.2	693	39.43	stable	alpha beta	lyase activity
YDR238C	0.48	0.427	112.412.178	44.07	4.994	1.064.962.726	0.218	11	1.003	YES	4.2	973	39.92	stable	all alpha	molecular function unknown
YDR243C	0.415	0.601	0.690515807	8.01	2.985	0.701262272	0.162	5	1.005	YES	0	588	37.54	stable	all alpha	Various
YDR244W	0.512	0.547	0.936014625	7.46	2.534	1.477.104.874	0.114	8	1.003	NO	2.5	612	44.36	unstable	all alpha	protein binding
YDR264C	0.5	0.505	0.99009901	93.14	3.54	0.796812749	0.133	10	0.985	NO	1.7	764	29.35	stable	Membrane	transferase activity
YDR280W	0.442	0.342	1.292.397.661	22.69	4.261	0.922509225	0.136	18	0.982	YES	0	305	51.83	unstable	alpha beta	hydrolase activity
YDR324C	0.459	0.593	0.774030354	56.53	4.191	0.9765625	0.169	8	1.001	YES	5	776	35.1	stable	all beta	RNA binding
YDR330W	0.483	0.571	0.845884413	6.82	3.073	3.267.973.856	0.16	1	0.994	NO	16.8	500	46.99	unstable	alpha beta	molecular function unknown
YDR339C	0.269	0.875	0.307428571	7.57	5.543	0.923361034	0.154	2	0.979	YES	0	189	22.58	stable	all alpha	molecular function unknown
YDR341C	0.291	0.773	0.376455369	591.24	5.047	0.926784059	0.285	3	0.952	YES	0	607	33.22	stable	all alpha	ligase activity
YDR346C	0.528	0.572	0.923076923	271.07	4.979	0.996015936	0.254	1	1	NO	11	481	49.6	unstable	all beta	molecular function unknown
YDR353W	0.219	0.673	0.325408618	307.51	5.988	0.941619586	0.315	10	0.97	YES	0	319	38.8	stable	alpha beta	oxidoreductase activity
YDR354W	0.487	0.463	1.051.835.853	87.48	4.649	0.039016777	0.142	0	1.018	NO	0	380	34.09	stable	alpha beta	transferase activity
YDR361C	0.453	0.587	0.771720613	39.94	6.727	0.684931507	0.212	2	0.988	YES	14.5	283	46.2	unstable	alpha beta	molecular function unknown
YDR385W	0.176	0.602	0.292358804	3624.33	4.599	1.102.535.832	0.8	6	0.939	NO	1.3	842	31.23	stable	alpha beta	translation regulator activity
YDR388W	0.387	0.659	0.587253414	169.74	4.81	1.443.001.443	0.177	72	0.947	NO	23.2	482	46.25	unstable	all alpha	protein binding
YDR404C	0.326	0.694	0.469740634	23.47	3.556	1.057.082.452	0.152	14	0.874	YES	0	171	37.4	stable	all beta	nucleotidyltransferase activity
YDR418W	0.206	0.622	0.331189711	1745.95	3.045	0.99009901	0.766	6	0.873	NO	15.8	165	33.54	stable	alpha beta	structural molecule activity
YDR429C	0.492	0.575	0.855652174	74.81	5.221	1.326.259.947	0.249	14	0.963	YES	5.5	274	51.12	unstable	alpha beta	translation regulator activity
YDR496C	0.457	0.573	0.797556719	74.07	5.605	0.995024876	0.247	18	0.984	NO	11	656	41.23	unstable	all alpha	transcription regulator activity
YDR502C	0.175	0.667	0.262368816	644.62	6.862	1.245.330.012	0.498	3	0.998	NO	6.5	384	35.35	stable	alpha beta	transferase activity
YEL013w	0.286	0.793	0.360655738	220.36	3.694	1.046.025.105	0.186	10	0.988	NO	3.8	578	44.84	unstable	all alpha	protein binding
YEL027w	0.177	0.633	0.279620853	769.3	4.835	0.915750916	0.584	11	0.987	NO	18.8	160	22.79	stable	Membrane	transporter activity
YEL037c	0.443	0.556	0.79676259	53.59	5.66	1.082.251.082	0.164	35	0.999	NO	31.2	398	47.69	unstable	all alpha	protein binding
YEL046c	0.494	0.411	1.201.946.472	922.37	6.202	1.081.081.081	0.33	3	0.98	NO	0	387	27.44	stable	alpha beta	lyase activity
YEL058w	0.421	0.563	0.747779751	87.1	4.763	1.057.082.452	0.156	1	0.987	YES	0	557	33.41	stable	alpha beta	isomerase activity
YEL060c	0.435	0.38	1.144.736.842	91.01	5.095	0.945179584	0.3	18	1.001	NO	13.4	635	32.36	stable	alpha beta	peptidase activity
YER006w	0.414	0.456	0.907894737	119.22	5.397	1.005.025.126	0.21	18	0.99	YES	13.7	520	46.12	unstable	all alpha	hydrolase activity
YER012w	0.332	0.765	0.433986928	14.55	5.916	0.877963126	0.204	36	0.988	YES	0	198	37.18	stable	all beta	hydrolase activity
YER021W	0.464	0.529	0.877126654	130.81	4.708	1.811.594.203	0.183	18	0.977	YES	4.4	523	42.94	unstable	all alpha	molecular function unknown

YER023w	0.5	0.516	0.968992248	143.35	6.167	1.020.408.163	0.21	8	1.003	YES	5.2	286	25.94	stable	alpha beta	oxidoreductase activity
YER025w	0.174	0.73	0.238356164	783.67	5.241	0.900900901	0.333	17	1.002	YES	7.6	527	45.78	unstable	alpha beta	translation regulator activity
YER036c	0.395	0.415	0.951807229	652.8	4.338	2.739.726.027	0.372	6	1.001	YES	5.9	610	41.64	unstable	alpha beta	hydrolase activity
YER043c	0.156	0.649	0.2403698	2328.52	6.119	0.947867299	0.641	13	1.013	YES	3.8	449	35.83	stable	alpha beta	hydrolase activity
YER055c	0.385	0.556	0.692446043	419.2	4.335	0.894454383	0.192	0	1.018	NO	0	297	24.65	stable	alpha beta	transferase activity
YER068w	0.454	0.528	0.859848485	29.84	1.732	1.615.508.885	0.151	9	1.011	NO	12.6	587	43.81	unstable	alpha beta	Various
YER069w	0.411	0.38	1.081.578.947	56.46	4.079	1.122.334.456	0.198	2	0.983	NO	1.3	863	31.66	stable	alpha beta	transferase activity
YER086w	0.431	0.322	1.338.509.317	70.82	3.421	0.888099467	0.312	10	0.995	NO	6.1	576	37.14	stable	alpha beta	lyase activity
YER089c	0.437	0.614	0.711726384	138.16	4.741	1.131.221.719	0.142	2	1.006	NO	4.3	464	40.41	unstable	alpha beta	phosphoprotein phosphatase
YER090w	0.278	0.687	0.404657933	171.58	4.675	0.750187547	0.216	4	0.985	NO	0	507	40.09	unstable	alpha beta	lyase activity
YER091c	0.356	0.484	0.73553719	1072.45	5.326	103.950.104	0.657	3	0.988	NO	2.2	767	33.76	stable	alpha beta	transferase activity
YER094c	0.41	0.29	1.413.793.103	28.79	4.467	1.057.082.452	0.159	9	0.991	YES	0	205	34.88	stable	all beta	peptidase activity
YER133w	0.14	0.8	0.175	93.97	4.854	0.871080139	0.229	60	0.971	YES	0	312	47.55	unstable	alpha beta	phosphoprotein phosphatase
YER136w	0.286	0.697	0.410329986	161.38	5.095	0.782472613	0.233	15	0.989	YES	0	451	41.02	unstable	alpha beta	enzyme regulator activity
YER148w	0.183	0.733	0.249658936	62.15	2.571	0.563380282	0.173	34	0.998	YES	0	240	36.19	stable	alpha beta	Various
YER156c	0.445	0.373	1.193.029.491	85.31	6.03	1.100.110.011	0.162	3	0.997	NO	3.6	338	38.75	stable	alpha beta	molecular function unknown
YER165w	0.43	0.388	1.108.247.423	32.76	2.984	0.871839582	0.488	20	0.981	YES	6.2	577	42.18	unstable	alpha beta	RNA binding
YER168c	0.485	0.344	1.409.883.721	34.15	4.284	0.507099391	0.137	2	0.994	YES	0	263	41.42	unstable	all alpha	transferase activity
YER178w	0.447	0.323	1.383.900.929	592.14	4.166	1.028.806.584	0.296	8	1.004	NO	0	420	39.08	stable	alpha beta	oxidoreductase activity
YFL002C	0.457	0.542	0.843173432	64.75	6.406	1.054.852.321	0.13	4	1.013	YES	10.7	606	36.55	stable	all alpha	helicase activity
YFL018C	0.367	0.351	1.045.584.046	70.2	5.407	1.034.126.163	0.253	12	0.962	NO	1.6	499	32.79	stable	alpha beta	oxidoreductase activity
YFL037W	0.179	0.771	0.232166018	496.9	5.083	0.780640125	0.271	6	0.974	YES	5.9	457	33.62	stable	alpha beta	structural molecule activity
YFL038C	0.217	0.848	0.255896226	30.26	6.951	1.005.025.126	0.185	19	0.988	YES	0	206	31.03	stable	alpha beta	hydrolase activity
YFL039C	0.088	0.648	0.135802469	2861.39	4.783	1.160.092.807	0.711	39	0.94	YES	0	375	40.04	unstable	alpha beta	structural molecule activity
YFL045C	0.358	0.343	1.043.731.778	671.85	7.567	0.762776506	0.54	14	0.985	YES	0	254	41.61	unstable	alpha beta	isomerase activity
YFR010W	0.514	0.42	1.223.809.524	54.27	5.228	0.807102502	0.208	14	1.008	NO	4	499	45.8	unstable	alpha beta	hydrolase activity
YFR037C	0.465	0.634	0.733438486	19.72	4.734	118.623.962	0.128	16	0.983	YES	8.8	557	41.82	unstable	alpha beta	hydrolase activity
YFR044C	0.414	0.383	1.080.939.948	352.41	6.884	1.119.820.829	0.313	2	1.013	NO	0	481	32.62	stable	alpha beta	hydrolase activity
YFR052W	0.468	0.46	1.017.391.304	10.91	5.15	1.512.859.304	0.18	26	1.002	YES	10.9	274	47.99	unstable	all alpha	peptidase activity
YGL008C	0.288	0.237	1.215.189.873	4400.34	5.111	1.040.582.726	0.734	8	0.946	YES	8	918	33.9	stable	Membrane	hydrolase activity
YGL009C	0.259	0.701	0.369472183	1121.57	4.609	1.107.419.712	0.336	1	0.977	NO	3.3	779	32.92	stable	alpha beta	lyase activity
YGL026C	0.346	0.437	0.791762014	337.07	5.405	1.094.091.904	0.32	5	1.001	NO	0	707	30.6	stable	alpha beta	lyase activity
YGL111W	0.524	0.618	0.84789644	31.69	3.651	0.788643533	0.127	10	1.017	YES	2.2	463	34.68	stable	alpha beta	molecular function unknown
YGL115W	0.47	0.414	11.352.657	32.32	4.088	1.808.318.264	0.16	24	1.007	NO	9	322	32.27	stable	alpha beta	Various
YGL120C	0.354	0.32	110.625	156.89	4.612	1.025.641.026	0.206	22	1.017	YES	3.1	767	44.81	unstable	alpha beta	Various
YGL135W	0.178	0.595	0.299159664	2598.21	2.948	1.082.251.082	0.832	3	1.019	NO	7.8	217	31.48	stable	alpha beta	structural molecule activity
YGL137W	0.437	0.422	1.035.545.024	360.39	4.442	0.893655049	0.208	66	1.011	YES	3.5	889	27.97	stable	alpha beta	molecular function unknown
YGL147C	0.318	0.579	0.549222798	931.62	5.923	0.911577028	0.771	1	0.999	NO	0	191	30.97	stable	alpha beta	structural molecule activity
YGL148W	0.234	0.748	0.312834225	181.96	4.692	1.008.064.516	0.323	1	1.003	NO	0	376	34.49	stable	alpha beta	oxidoreductase activity
YGL155W	0.52	0.538	0.966542751	20.03	2.205	1.049.317.943	0.111	1	1.006	YES	4.8	376	36.54	stable	all alpha	signal transducer activity
YGL157W	0.396	0.568	0.697183099	34.26	4.491	0.981354269	0.206	3	1.017	NO	0	347	25.88	stable	alpha beta	oxidoreductase activity
YGL171W	0.417	0.579	0.720207254	21.78	4.697	1.191.895.113	0.171	4	1.014	YES	4.6	564	44.81	unstable	all alpha	helicase activity
YGL201C	0.388	0.364	1.065.934.066	56.15	3.379	0.858369099	0.172	9	1.018	YES	8.7	1017	48.06	unstable	alpha beta	helicase activity
YGL221C	0.403	0.553	0.72875226	12.26	4.629	0.797448166	0.182	3	1.015	NO	0	288	23.85	stable	alpha beta	molecular function unknown
YGL244W	0.483	0.645	0.748837209	20.92	4.567	0.843881857	0.198	9	1.009	NO	16.1	558	47.3	unstable	all alpha	transcription regulator activity
YGL245W	0.308	0.706	0.436260623	328.67	4.766	0.757575758	0.461	16	1.01	YES	8.1	708	31.45	stable	alpha beta	ligase activity
YGL253W	0.453	0.363	1.247.933.884	2136.15	5.858	1.858.736.059	0.643	1	0.74	NO	2.7	486	39.75	stable	alpha beta	transferase activity

YGR007W	0.426	0.572	0.744755245	27.14	2.865	0.773993808	0.143	0	1.014	NO	0	323	28.06	stable	alpha beta	nucleotidyltransferase activity
YGR019W	0.45	0.36	1.25	15.69	4.667	0.759301443	0.287	0	1.013	NO	4.2	471	33.53	stable	alpha beta	transferase activity
YGR054W	0.433	0.546	0.793040293	288.13	4.501	0.805152979	0.219	5	1.02	NO	10.6	642	47.08	unstable	all beta	translation regulator activity
YGR061C	0.358	0.492	0.727642276	154.07	3.311	2.141.327.623	0.277	2	1.014	NO	3.2	1358	38.1	stable	alpha beta	ligase activity
YGR078C	0.328	0.767	0.427640156	6.4	5.968	0.902527076	0.123	5	1.013	NO	14.6	199	40.38	unstable	all alpha	protein binding
YGR090W	0.494	0.547	0.903107861	50.56	3.759	1.189.060.642	0.187	51	0.995	YES	6.1	1237	36.84	stable	alpha beta	RNA binding
YGR094W	0.278	0.695	0.4	1422.98	3.456	0.910746812	0.369	4	1.016	YES	5.8	1104	37.59	stable	alpha beta	ligase activity
YGR118W	0.136	0.65	0.209230769	761.13	5.18	1.061.571.125	0.726	0	1.01	NO	0	145	23.84	stable	alpha beta	structural molecule activity
YGR123C	0.47	0.313	1.501.597.444	120.99	4.563	0.868809731	0.173	12	0.982	NO	2.3	513	31.24	stable	all alpha	hydrolase activity
YGR124W	0.322	0.329	0.978723404	848.34	5.297	1.416.430.595	0.317	2	0.985	NO	4.9	572	36.95	stable	alpha beta	ligase activity
YGR173W	0.285	0.705	0.404255319	22.13	6.028	1.088.139.282	0.206	6	0.981	NO	2.4	368	34.8	stable	alpha beta	molecular function unknown
YGR175C	0.472	0.448	1.053.571.429	917.65	5.513	0.897666068	0.441	4	0.994	YES	0	496	32.59	stable	Membrane	oxidoreductase activity
YGR187C	0.496	0.335	1.480.597.015	35.41	5.707	1.004.016.064	0.184	5	0.975	NO	4.8	394	48.97	unstable	all alpha	molecular function unknown
YGR207C	0.343	0.783	0.438058748	5.2	6.274	1.096.491.228	0.178	2	0.764	NO	0	261	38.92	stable	alpha beta	molecular function unknown
YGR211W	0.439	0.352	1.247.159.091	180.49	5.437	1.207.729.469	0.244	4	0.981	YES	2.7	486	43.46	unstable	alpha beta	protein binding
YGR218W	0.36	0.501	0.718562874	232.61	3.601	0.670241287	0.205	64	0.957	YES	3	1084	41.52	unstable	all alpha	protein binding
YGR232W	0.479	0.531	0.902071563	5.71	4.727	0.833333333	0.167	8	1.011	NO	0	228	27.01	stable	all alpha	molecular function unknown
YGR234W	0.442	0.632	0.699367089	1505.37	7.245	0.997008973	0.267	11	1.017	NO	4	399	37.02	stable	alpha beta	oxidoreductase activity
YGR253C	0.25	0.704	0.355113636	27.83	5.528	0.928505107	0.162	7	0.989	YES	8.5	260	49.38	unstable	alpha beta	peptidase activity
YGR260W	0.523	0.442	1.183.257.919	349.66	3.494	0.547645126	0.193	38	0.991	NO	2.2	534	33.09	stable	Membrane	transporter activity
YGR264C	0.318	0.653	0.486983155	74.24	4.945	1.170.960.187	0.293	4	0.998	YES	0	751	41.72	unstable	alpha beta	ligase activity
YGR285C	0.357	0.648	0.550925926	507.83	6.162	0.871839582	0.504	11	0.993	NO	17.1	433	39.8	stable	all alpha	protein binding
YHR019c	0.414	0.375	1.104	922.62	4.891	0.975609756	0.4	7	1.01	YES	7.2	554	40.42	unstable	alpha beta	ligase activity
YHR020W	0.414	0.33	1.254.545.455	710.26	4.83	0.939849624	0.355	6	1.01	YES	4.2	688	44.22	unstable	alpha beta	ligase activity
YHR025w	0.418	0.398	1.050.251.256	190.34	5.025	1.414.427.157	0.271	2	1.016	NO	5.3	357	44.11	unstable	alpha beta	transferase activity
YHR030c	0.401	0.407	0.985257985	62.85	3.665	1.113.585.746	0.138	46	1.02	NO	6.2	484	46.22	unstable	alpha beta	Various
YHR042w	0.481	0.471	1.021.231.423	406.95	4.116	1.221.001.221	0.226	7	1.015	YES	4.8	691	33.85	stable	alpha beta	oxidoreductase activity
YHR051w	0.369	0.757	0.487450462	40.52	5.304	0.319284802	0.254	2	0.99	NO	0	148	52.51	unstable	all alpha	transporter activity
YHR064C	0.424	0.652	0.650306748	254.44	5.339	0.255819903	0.455	12	0.98	NO	3.3	538	28.48	stable	alpha beta	protein binding
YHR068W	0.401	0.322	1.245.341.615	169.11	5.293	1.063.829.787	0.419	3	0.989	YES	2.1	387	33.55	stable	alpha beta	transferase activity
YHR072w	0.393	0.737	0.533242877	150.57	3.462	0.81300813	0.147	1	0.998	YES	1.9	731	39.27	stable	all alpha	isomerase activity
YHR074W	0.334	0.619	0.539579968	41.49	4.767	2.577.319.588	0.172	2	0.988	YES	1.3	714	46.25	unstable	alpha beta	ligase activity
YHR112C	0.427	0.626	0.682108626	16.4	4.126	1.329.787.234	0.178	5	0.99	NO	0	378	43.37	unstable	alpha beta	lyase activity
YHR170w	0.401	0.301	1.332.225.914	251.83	4.113	2.277.904.328	0.244	14	0.988	YES	5	518	37.2	stable	alpha beta	RNA binding
YHR183w	0.234	0.496	0.471774194	1182.23	6.408	1.089.324.619	0.623	15	0.995	NO	2.2	489	33.63	stable	all alpha	oxidoreductase activity
YIL020C	0.304	0.691	0.439942113	10.55	4.527	0.743494424	0.161	1	0.988	NO	0	261	23.34	stable	alpha beta	isomerase activity
YIL021W	0.307	0.78	0.393589744	11.1	5.108	1.058.201.058	0.167	19	0.941	YES	0	318	33.7	stable	alpha beta	nucleotidyltransferase activity
YIL030C	0.533	0.63	0.846031746	50.52	3.258	1.022.494.888	0.169	1	0.989	NO	7.1	1319	38.43	stable	Membrane	ligase activity
YIL033C	0.51	0.449	1.135.857.461	87.64	5.504	1.430.615.165	0.178	15	0.824	NO	10.8	416	52.11	unstable	alpha beta	enzyme regulator activity
YIL063C	0.512	0.594	0.861952862	5.74	4.107	1.096.491.228	0.17	8	0.996	YES	14.4	327	39.19	stable	all beta	molecular function unknown
YIL075C	0.423	0.525	0.805714286	513.46	3.798	0.590667454	0.176	5	0.975	YES	5.3	945	33.49	stable	alpha beta	Various
YIL078W	0.409	0.22	1.859.090.909	31.23	3.938	0.786782061	0.408	3	0.982	YES	1.9	734	42.58	unstable	alpha beta	ligase activity
YIL109C	0.437	0.629	0.694753577	474.58	3.471	117.370.892	0.212	11	0.938	YES	7.8	926	56	unstable	alpha beta	protein binding
YIL116W	0.403	0.489	0.824130879	27.37	3.628	0.860585198	0.209	0	0.934	NO	0	385	27.19	stable	alpha beta	transferase activity
YIL118W	0.353	0.317	1.113.564.669	30.56	3.699	0.770416025	0.182	8	0.964	YES	8.2	231	43	unstable	alpha beta	signal transducer activity
YIL142W	0.429	0.185	2.318.918.919	107.4	5.631	0.921658986	0.193	25	0.897	YES	4.9	527	35	stable	alpha beta	protein binding
YIL145C	0.418	0.75	0.557333333	39.59	3.503	8.547.008.547	0.126	2	0.96	NO	0	309	34.65	stable	alpha beta	ligase activity

YIR008C	0.422	0.622	0.678456592	12.16	4.547	1.261.034.048	0.164	5	0.973	YES	3.2	409	43.28	unstable	Multidomain	nucleotidyltransferase activity
YIR026C	0.505	0.606	0.833333333	23.9	5.38	118.623.962	0.164	2	0.985	NO	4.1	364	41.62	unstable	alpha beta	phosphoprotein phosphatase
YIR034C	0.372	0.503	0.739562624	46.52	6.533	0.931098696	0.218	6	0.979	NO	3.5	373	34.71	stable	alpha beta	oxidoreductase activity
YJL001W	0.258	0.745	0.346308725	45.01	5.556	1.633.986.928	0.172	10	0.982	YES	0	215	14.93	stable	all beta	peptidase activity
YJL014W	0.332	0.348	0.954022989	270.5	5.887	1.003.009.027	0.228	16	0.992	YES	0	534	46.15	unstable	alpha beta	protein binding
YJL026W	0.368	0.317	1.160.883.281	532.45	6.334	1.098.901.099	0.501	16	0.767	YES	3.5	399	36.48	stable	all alpha	oxidoreductase activity
YJL050W	0.298	0.602	0.495016611	68.87	3.509	0.796178344	0.204	4	0.99	YES	6.5	1073	39.94	stable	alpha beta	helicase activity
YJL111W	0.383	0.251	1.525.896.414	230.75	4.909	1.308.900.524	0.192	1	0.981	YES	0	550	29.04	stable	alpha beta	protein binding
YJL140W	0.546	0.383	1.425.587.467	11.69	3.861	1.107.419.712	0.137	11	0.914	NO	21.3	221	48.88	unstable	all alpha	transferase activity
YJL167W	0.3	0.753	0.398406375	95.92	7.276	1.175.088.132	0.373	6	1.021	YES	0	352	36.8	stable	all alpha	transferase activity
YJL172W	0.468	0.496	0.943548387	151.03	6.306	0.697836706	0.25	3	1.001	NO	2.1	576	35.4	stable	Membrane	hydrolase activity
YJL200C	0.33	0.549	0.601092896	81.23	4.521	1.075.268.817	0.219	1	1.007	NO	2.9	789	27.67	stable	alpha beta	lyase activity
YJR002W	0.453	0.647	0.70015456	49.92	5.926	0.859106529	0.169	8	0.984	YES	15.9	593	58.51	unstable	all alpha	molecular function unknown
YJR007W	0.265	0.763	0.347313237	120.62	5.734	0.985221675	0.371	16	0.99	YES	8.9	304	52.56	unstable	alpha beta	translation regulator activity
YJR016C	0.388	0.301	1.289.036.545	1258.33	3.861	1.102.535.832	0.378	1	1	YES	1.7	585	34.73	stable	alpha beta	lyase activity
YJR024C	0.478	0.509	0.939096267	39.54	2.441	2.336.448.598	0.121	1	0.993	NO	0	244	46.77	unstable	alpha beta	molecular function unknown
YJR064W	0.235	0.751	0.312916112	270.8	4.222	0.977517107	0.217	15	0.994	YES	0	562	30.26	stable	alpha beta	protein binding
YJR104C	0.281	0.695	0.404316547	84.19	6.098	0.999000999	0.377	9	1.002	NO	0	154	24.8	stable	all beta	oxidoreductase activity
YJR109C	0.265	0.661	0.400907716	299.98	4.215	1.097.694.841	0.239	8	1.01	NO	4.1	1118	35.68	stable	alpha beta	ligase activity
YJR144W	0.492	0.42	1.171.428.571	5.88	4.257	0.786163522	0.163	10	0.996	NO	5.9	269	26.85	stable	alpha beta	DNA binding
YJR148W	0.441	0.272	1.621.323.529	73.52	5.053	0.385208012	0.195	3	1.003	NO	0	376	27.23	stable	Multidomain	transferase activity
YKL007W	0.443	0.675	0.656296296	12.72	5.187	0.972762646	0.181	8	1.006	NO	3.7	268	46.09	unstable	alpha beta	protein binding
YKL009W	0.392	0.666	0.588588589	47.37	5.124	0.857632933	0.279	7	0.786	NO	0	236	42.53	unstable	alpha beta	molecular function unknown
YKL021C	0.512	0.57	0.898245614	36.24	4.923	1.153.402.537	0.193	5	1.004	YES	10.9	468	34.6	stable	all beta	molecular function unknown
YKL035W	0.337	0.452	0.745575221	250.45	6.009	1.107.419.712	0.33	1	1.005	YES	2.2	499	31.06	stable	alpha beta	transferase activity
YKL060C	0.245	0.6	0.408333333	4286.55	6.605	1.137.656.428	0.869	4	1.015	YES	0	359	32.11	stable	alpha beta	lyase activity
YKL081W	0.393	0.679	0.578792342	2693.38	5.688	1.020.408.163	0.553	22	1.03	NO	9	412	35.83	stable	alpha beta	translation regulator activity
YKL113C	0.292	0.831	0.351383875	28.78	7.011	1.196.172.249	0.16	5	1.006	NO	6.5	382	40.31	unstable	all alpha	hydrolase activity
YKL120W	0.403	0.386	1.044.041.451	88.32	7.587	0.044881289	0.187	6	1.004	NO	4.3	324	36.36	stable	Membrane	transporter activity
YKL145W	0.157	0.687	0.22852984	78.82	5.43	2.032.520.325	0.232	32	1.003	YES	10.9	467	37.35	stable	all alpha	peptidase activity
YKL148C	0.358	0.274	1.306.569.343	101.02	5.061	110.864.745	0.245	1	1.007	NO	4.8	640	38.38	stable	alpha beta	oxidoreductase activity
YKL181W	0.389	0.409	0.951100244	320.19	3.491	1.138.952.164	0.255	3	1.008	NO	4.2	427	41.51	unstable	alpha beta	transferase activity
YKL182W	0.455	0.351	1.296.296.296	912.07	2.473	1.404.494.382	0.364	1	1.012	YES	1.5	2051	33.07	stable	alpha beta	Various
YKL195W	0.438	0.546	0.802197802	9.22	4.294	0.821692687	0.183	2	0.992	YES	15.9	403	60.59	unstable	Membrane	molecular function unknown
YKL196C	0.301	0.829	0.363088058	13.73	4.675	131.061.599	0.181	11	0.999	YES	0	200	46.71	unstable	alpha beta	transferase activity
YKL209C	0.506	0.602	0.840531561	82.45	3.109	2.070.393.375	0.127	3	0.99	NO	2.8	1290	35.25	stable	Membrane	hydrolase activity
YKL210W	0.361	0.494	0.730769231	198.72	4.093	0.960614793	0.212	15	0.971	YES	1.3	1024	25.62	stable	alpha beta	Other
YKL211C	0.316	0.746	0.423592493	108.99	5.624	1.027.749.229	0.184	5	1.001	NO	0	484	41.82	unstable	alpha beta	lyase activity
YKL216W	0.539	0.529	1.018.903.592	72.43	7.38	1.597.444.089	0.225	2	0.968	NO	0	314	23.73	stable	alpha beta	oxidoreductase activity
YKR048C	0.383	0.778	0.492287918	54.11	6.259	1.057.082.452	0.153	28	1.001	NO	12.2	417	54.19	unstable	alpha beta	protein binding
YLL008w	0.399	0.54	0.738888889	92.09	4.331	0.761614623	0.227	8	0.985	YES	13.7	752	48.93	unstable	all alpha	hydrolase activity
YLL018c	0.386	0.331	1.166.163.142	808.5	5.537	1.098.901.099	0.35	4	0.985	YES	8.1	557	46.07	unstable	alpha beta	RNA binding
YLL031c	0.428	0.6	0.713333333	51.82	3.634	0.236910685	0.165	1	0.966	YES	6.6	1017	30.05	stable	Membrane	transferase activity
YLL034c	0.425	0.321	1.323.987.539	55.61	3.143	1.121.076.233	0.173	6	0.989	YES	5.1	837	46.84	unstable	all alpha	hydrolase activity
YLR027c	0.444	0.417	1.064.748.201	507.65	4.435	1.009.081.736	0.232	4	0.995	NO	0	418	31.16	stable	alpha beta	transferase activity
YLR058c	0.253	0.605	0.418181818	135.27	6.648	0.604229607	0.589	7	0.992	NO	2.6	469	27.27	stable	all alpha	transferase activity
YLR059c	0.433	0.477	0.907756813	14.39	2.945	0.964320154	0.144	6	1.013	NO	6.3	269	50.86	unstable	alpha beta	hydrolase activity

YLR060w	0.36	0.6	0.6	466.07	4.951	0.952380952	0.325	1	0.904	YES	0	595	44.66	unstable	alpha beta	ligase activity
YLR109w	0.441	0.582	0.757731959	124.85	8.609	0.987166831	0.549	11	0.998	NO	0	176	34.33	stable	alpha beta	oxidoreductase activity
YLR113w	0.144	0.813	0.177121771	179.33	3.574	1.251.564.456	0.175	10	0.977	NO	4.8	435	28.5	stable	alpha beta	Various
YLR153c	0.425	0.308	137.987.013	923.01	5.799	134.589.502	0.371	5	1.007	YES	0	683	32.03	stable	alpha beta	ligase activity
YLR163C	0.481	0.277	1.736.462.094	20.01	4.045	0.983284169	0.143	7	1.028	YES	2.4	462	36.95	stable	alpha beta	peptidase activity
YLR167W	0.14	0.619	0.226171244	545.94	3.106	1.024.590.164	0.811	1	1.006	YES	17.8	152	27.42	stable	alpha beta	structural molecule activity
YLR175W	0.342	0.221	1.547.511.312	671.29	4.718	1.189.060.642	0.375	28	1.021	YES	21.3	483	44.03	unstable	alpha beta	isomerase activity
YLR186W	0.294	0.734	0.400544959	5.24	5.41	1.138.952.164	0.206	9	1.013	YES	0	252	40.31	unstable	alpha beta	RNA binding
YLR196W	0.376	0.683	0.550512445	122.85	4.882	0.856164384	0.239	25	1.017	YES	7.1	576	39.83	stable	all beta	molecular function unknown
YLR197W	0.289	0.74	0.390540541	698.15	5.122	0.843881857	0.37	16	0.836	YES	11.9	504	32.56	stable	all alpha	molecular function unknown
YLR216C	0.37	0.659	0.561456753	50.36	6.369	0.838222967	0.253	25	0.998	NO	6.5	371	26.89	stable	alpha beta	isomerase activity
YLR244C	0.422	0.303	1.392.739.274	75.84	4.814	1.282.051.282	0.291	0	0.947	NO	2.1	387	39.86	stable	alpha beta	hydrolase activity
YLR259C	0.294	0.532	0.552631579	553.87	7.692	1.335.113.485	0.382	1	0.981	YES	7	572	39.52	stable	alpha beta	DNA binding
YLR276C	0.413	0.423	0.976359338	54.28	3.924	1.248.439.451	0.18	7	0.974	YES	7.4	594	39.77	stable	all alpha	hydrolase activity
YLR293C	0.098	0.689	0.142235123	283.56	7.13	0.683060109	0.621	26	0.889	NO	13.2	219	33.75	stable	alpha beta	hydrolase activity
YLR300W	0.381	0.68	0.560294118	784.59	6.653	1.254.705.144	0.345	2	0.983	NO	3.8	448	35.22	stable	alpha beta	hydrolase activity
YLR304C	0.349	0.217	1.608.294.931	985.77	4.515	1.103.752.759	0.462	15	0.975	NO	3.5	778	25.73	stable	alpha beta	lyase activity
YLR314C	0.451	0.403	11.191.067	23.67	5.064	0.937207123	0.183	16	0.989	YES	4.8	520	50.19	unstable	alpha beta	structural molecule activity
YLR347C	0.413	0.636	0.649371069	156.14	3.905	0.859106529	0.195	57	0.995	YES	1.7	861	44.75	unstable	all alpha	transporter activity
YLR351C	0.341	0.684	0.498538012	18.18	2.526	0.73964497	0.121	1	0.989	NO	5.8	291	38.92	stable	alpha beta	hydrolase activity
YLR355C	0.237	0.584	0.405821918	3020.14	7.154	1.024.590.164	0.802	6	1	YES	5.3	395	30.27	stable	alpha beta	oxidoreductase activity
YLR370C	0.36	0.6	0.6	6.81	6.015	1.081.081.081	0.158	9	1.003	NO	0	178	39.7	stable	all alpha	structural molecule activity
YLR380W	0.508	0.491	1.034.623.218	54.79	6.293	0.947867299	0.245	0	0.98	NO	2.9	408	46.64	unstable	all alpha	transporter activity
YLR384C	0.449	0.526	0.853612167	31.7	3.033	1.253.132.832	0.181	12	0.989	NO	4.7	1349	39.73	stable	alpha beta	molecular function unknown
YLR398C	0.355	0.672	0.52827381	22.19	4.026	1.218.026.797	0.178	4	0.99	NO	7.7	1287	42.45	unstable	alpha beta	translation regulator activity
YLR409C	0.424	0.584	0.726027397	79.39	3.271	0.562746201	0.176	8	0.98	YES	4.4	939	42.08	unstable	alpha beta	RNA binding
YLR410W	0.319	0.666	0.478978979	81.75	3.183	1.122.334.456	0.164	0	0.972	NO	7.2	1146	44.55	unstable	alpha beta	transferase activity
YLR420W	0.428	0.404	1.059.405.941	24.4	5.061	1.182.033.097	0.149	2	1.017	NO	0	364	35.67	stable	alpha beta	hydrolase activity
YLR427W	0.525	0.54	0.972222222	36.91	1.863	1.483.679.525	0.141	26	1.003	NO	13	670	53.29	unstable	all alpha	molecular function unknown
YLR432W	0.274	0.694	0.39481268	1534.66	4.799	1.116.071.429	0.464	13	1.016	NO	2.3	523	31.34	stable	alpha beta	oxidoreductase activity
YLR447C	0.386	0.537	0.718808194	339.55	5.172	0.993048659	0.248	95	1.008	NO	3.8	345	42.14	unstable	all alpha	transporter activity
YML008C	0.333	0.661	0.503782148	149.69	7.148	0.924214418	0.308	9	1.008	NO	6	383	30.1	stable	all alpha	transferase activity
YML028W	0.256	0.714	0.358543417	170.11	8.27	110.864.745	0.714	3	0.988	NO	0	196	29.98	stable	Membrane	oxidoreductase activity
YML035C	0.329	0.737	0.446404342	79.39	3.112	0.760456274	0.19	2	1.002	NO	4	810	38.79	stable	alpha beta	hydrolase activity
YML063W	0.29	0.708	0.40960452	3.39	3.837	0.370233247	0.769	0	0.995	NO	11.8	255	33.21	stable	alpha beta	structural molecule activity
YML070W	0.425	0.498	0.853413655	59.33	6.254	1.169.590.643	0.217	1	1	NO	4.1	584	27.4	stable	alpha beta	transferase activity
YML080W	0.344	0.751	0.458055925	11.94	5.078	0.621118012	0.168	0	1.022	NO	4.3	423	35.11	stable	alpha beta	Other
YML085C	0.195	0.758	0.257255937	165.27	5.61	1.295.336.788	0.277	8	0.982	YES	5.6	447	31.48	stable	alpha beta	structural molecule activity
YML086C	0.497	0.462	1.075.757.576	160.27	5.826	0.951474786	0.217	0	0.997	NO	0	526	24.48	stable	alpha beta	oxidoreductase activity
YML094W	0.441	0.506	0.871541502	2.92	7.571	1.239.157.373	0.183	4	0.999	NO	5.5	163	47.56	unstable	all alpha	protein binding
YML105C	0.551	0.615	0.895934959	1086.07	5.009	0.578368999	0.173	3	0.998	YES	10.6	273	43.74	unstable	all alpha	molecular function unknown
YML123C	0.462	0.517	0.893617021	3782.96	5.642	0.754147813	0.461	28	0.972	NO	1.7	587	36.01	stable	Membrane	transporter activity
YML126C	0.279	0.739	0.377537212	1020.17	6.288	1.086.956.522	0.401	8	0.989	YES	0	491	32.09	stable	alpha beta	transferase activity
YML130C	0.488	0.52	0.938461538	102.93	5.368	0.946969697	0.217	7	0.997	YES	0	563	33.96	stable	all alpha	oxidoreductase activity
YMR011W	0.392	0.555	0.706306306	890.7	7.622	1.132.502.831	0.359	2	1.012	NO	2	541	30.46	stable	Membrane	transporter activity
YMR038C	0.563	0.596	0.944630872	4.35	5.493	0.962463908	0.158	3	1.003	NO	0	249	43.02	unstable	all beta	transporter activity
YMR079W	0.292	0.725	0.402758621	504.46	5.612	1.059.322.034	0.3	1	0.974	YES	5.3	304	45.24	unstable	all alpha	transporter activity

YMR093W	0.493	0.551	0.894736842	74.41	4.758	0.933706816	0.148	12	0.986	YES	0	513	42.27	unstable	alpha beta	RNA binding
YMR116C	0.38	0.452	0.840707965	2257.33	5.101	0.823723229	0.777	20	1.01	NO	0	319	25.36	stable	all beta	signal transducer activity
YMR146C	0.408	0.366	1.114.754.098	88.9	5.368	1.317.523.057	0.295	11	1.003	YES	0	347	28.51	stable	all beta	translation regulator activity
YMR203W	0.387	0.646	0.599071207	505.69	6.021	0.697836706	0.274	7	1	YES	0	387	35.05	stable	alpha beta	transporter activity
YMR205C	0.341	0.568	0.600352113	439.98	4.087	1.098.901.099	0.512	8	0.994	NO	4.4	959	32.75	stable	alpha beta	transferase activity
YMR217W	0.224	0.738	0.303523035	794.44	4.989	1.101.321.586	0.462	0	1.004	NO	0	525	27.82	stable	alpha beta	ligase activity
YMR229C	0.463	0.561	0.825311943	3992.84	0.361	147.275.405	0.237	9	1.006	YES	9.2	1729	39.79	stable	alpha beta	RNA binding
YMR235C	0.454	0.617	0.735818476	126.72	3.828	1.175.088.132	0.218	11	0.996	YES	10.8	407	35.6	stable	alpha beta	enzyme regulator activity
YMR290C	0.315	0.301	1.046.511.628	99.11	5.376	1	0.225	25	1.02	YES	5.3	505	37.62	stable	all alpha	Various
YMR297W	0.426	0.421	1.011.876.485	351.07	5.245	1.009.081.736	0.257	3	1.014	NO	7.5	532	39.49	stable	alpha beta	Various
YMR314W	0.405	0.394	1.027.918.782	30.48	4.966	1.009.081.736	0.159	14	0.975	YES	0	234	26.33	stable	alpha beta	hydrolase activity
YMR315W	0.479	0.528	0.90719697	52.6	6.132	1.324.503.311	0.253	3	0.988	NO	0	349	22.38	stable	alpha beta	molecular function unknown
YNL001W	0.454	0.553	0.820976492	9.64	4.987	0.815660685	0.151	0	0.969	NO	3.1	386	32.82	stable	alpha beta	molecular function unknown
YNL021W	0.41	0.587	0.69846678	28.67	4.677	0.44345898	0.151	8	0.981	NO	5	706	41.78	unstable	alpha beta	hydrolase activity
YNL024C	0.51	0.614	0.830618893	3.82	2.711	0.239463602	0.147	2	0.938	NO	6.1	246	26.31	stable	alpha beta	transferase activity
YNL055C	0.455	0.585	0.777777778	484.48	8.01	0.937207123	0.361	19	0.993	NO	0	283	34.24	stable	alpha beta	transporter activity
YNL061W	0.353	0.497	0.710261569	126.67	4.972	1.369.863.014	0.254	48	1.004	YES	18.1	618	44.16	unstable	all alpha	transferase activity
YNL064C	0.33	0.705	0.468085106	476.49	5.6	0.931098696	0.373	7	0.971	NO	17.8	409	34.45	stable	alpha beta	protein binding
YNL104C	0.356	0.645	0.551937984	740.47	4.307	1.076.426.265	0.292	1	1.022	NO	2.1	619	38.34	stable	alpha beta	transferase activity
YNL113W	0.28	0.692	0.404624277	4.85	5.979	1.166.861.144	0.186	15	0.977	YES	4.2	142	52.2	unstable	alpha beta	transferase activity
YNL121C	0.504	0.604	0.834437086	95.4	5.917	1.129.943.503	0.274	10	0.994	NO	6.8	617	41.25	unstable	Membrane	transporter activity
YNL123W	0.418	0.487	0.858316222	88.87	3.496	0.914076782	0.178	0	0.976	NO	2.5	997	35.07	stable	all beta	nucleotidyltransferase activity
YNL142W	0.382	0.625	0.6112	77.31	4.918	0.46641791	0.213	1	1.005	NO	2.2	499	27.52	stable	Membrane	transporter activity
YNL163C	0.364	0.609	0.597701149	13.13	3.895	0.456621005	0.151	1	1.015	YES	2.3	1110	48.98	unstable	alpha beta	hydrolase activity
YNL182C	0.508	0.56	0.907142857	32.14	4.046	1.254.705.144	0.148	12	0.996	YES	6.1	555	34.28	stable	alpha beta	molecular function unknown
YNL189W	0.372	0.407	0.914004914	186.04	5.284	0.865800866	0.271	197	0.997	YES	4.1	542	43.36	unstable	all alpha	protein binding
YNL192W	0.428	0.5	0.856	64.87	1.929	1.182.033.097	0.15	3	1.001	NO	3.7	1131	45.68	unstable	Membrane	transferase activity
YNL219C	0.471	0.515	0.914563107	137.92	6.108	1.051.524.711	0.156	1	1.002	NO	4.5	555	39.3	stable	Membrane	transferase activity
YNL232W	0.499	0.432	1.155.092.593	11.93	5.35	0.958772771	0.119	12	0.66	YES	11.3	292	45.39	unstable	alpha beta	hydrolase activity
YNL241C	0.44	0.45	0.977777778	549.4	5.285	0.834724541	0.172	1	1.001	NO	0	505	33.7	stable	alpha beta	oxidoreductase activity
YNL287W	0.46	0.547	0.84095064	210.42	3.638	1.479.289.941	0.238	16	0.995	YES	2.7	935	40.86	unstable	alpha beta	molecular function unknown
YNL290W	0.352	0.741	0.475033738	10.34	4.607	0.438596491	0.12	17	0.981	YES	0	340	38.49	stable	all alpha	DNA binding
YNL297C	0.539	0.57	0.945614035	406.36	2.109	3.717.472.119	0.133	0	0.977	NO	6.1	1636	44.85	unstable	alpha beta	enzyme regulator activity
YNL301C	0.295	0.711	0.414908579	1617.89	4.546	1.051.524.711	0.68	0	0.975	NO	14	186	40.14	unstable	alpha beta	structural molecule activity
YNL313C	0.495	0.563	0.879218472	31.54	4.392	2.403.846.154	0.158	7	0.98	YES	1.1	904	43.1	unstable	all alpha	molecular function unknown
YNR003C	0.416	0.673	0.618127786	5.2	5.813	1.213.592.233	0.141	15	0.922	YES	0	317	38.77	stable	all alpha	nucleotidyltransferase activity
YNR012W	0.463	0.531	0.871939736	42.8	3.235	0.935453695	0.168	10	0.968	NO	4.6	501	37.55	stable	alpha beta	transferase activity
YNR015W	0.457	0.589	0.775891341	14.76	2.651	0.918273646	0.14	2	0.977	NO	0	384	34.26	stable	alpha beta	Other
YNR016C	0.362	0.453	0.799116998	113.98	2.181	1.067.235.859	0.328	9	0.989	YES	2.3	2233	41.23	unstable	alpha beta	ligase activity
YNR033W	0.458	0.568	0.806338028	51.36	3.669	119.047.619	0.136	1	1.002	NO	1.3	787	40.07	unstable	alpha beta	ligase activity
YNR036C	0.336	0.824	0.40776699	13.9	2.369	0.777000777	0.093	1	0.977	NO	0	153	32.35	stable	alpha beta	structural molecule activity
YNR043W	0.376	0.438	0.858447489	360.29	6.052	1.945.525.292	0.2	1	0.801	YES	4.8	396	39.84	stable	alpha beta	lyase activity
YNR046W	0.368	0.84	0.438095238	8.84	7.415	0.818330606	0.173	5	0.982	YES	0	135	39.82	stable	all alpha	transferase activity
YNR050C	0.336	0.568	0.591549296	130.13	5.327	1.152.073.733	0.332	5	0.996	NO	0	446	26.43	stable	alpha beta	oxidoreductase activity
YNR053C	0.344	0.445	0.773033708	312.26	3.466	0.323729362	0.219	17	1.002	YES	3.5	486	39.94	stable	alpha beta	hydrolase activity
YNR054C	0.453	0.629	0.720190779	8.79	5.576	1.184.834.123	0.179	2	0.994	YES	25	316	54.54	unstable	all alpha	transcription regulator activity
YOL010W	0.429	0.542	0.791512915	25.27	4.868	0.975609756	0.162	5	1.018	YES	0	367	40.33	unstable	alpha beta	molecular function unknown

YOL021C	0.443	0.317	1.397.476.341	29.7	3.95	1.019.367.992	0.178	17	0.993	YES	2.6	1001	40.11	unstable	alpha beta	hydrolase activity
YOL022C	0.511	0.584	0.875	26.73	4.784	1.079.913.607	0.164	1	0.991	YES	0	408	48.47	unstable	alpha beta	molecular function unknown
YOL030W	0.381	0.719	0.529902643	300.58	5.333	0.815660685	0.271	6	0.733	NO	18.8	484	42.84	unstable	alpha beta	transferase activity
YOL038W	0.316	0.757	0.417437252	33.64	4.645	7.299.270.073	0.156	10	0.976	YES	5.9	254	54.93	unstable	alpha beta	peptidase activity
YOL058W	0.292	0.628	0.464968153	197.87	6.87	1.180.637.544	0.403	3	1.009	NO	0	420	27.87	stable	alpha beta	ligase activity
YOL097C	0.323	0.544	0.59375	151.72	5.858	0.733675715	0.286	1	1.004	YES	0	432	34.88	stable	alpha beta	ligase activity
YOL098C	0.496	0.56	0.885714286	56.7	2.984	1.082.251.082	0.191	1	0.988	NO	3.1	1037	39.91	stable	alpha beta	molecular function unknown
YOL124C	0.519	0.448	1.158.482.143	18.8	4.487	0.314169023	0.142	1	1.016	NO	2.1	433	42.25	unstable	alpha beta	RNA binding
YOR007C	0.413	0.579	0.713298791	163.89	5.611	0.288600289	0.192	6	0.994	NO	6.1	346	47.8	unstable	all alpha	molecular function unknown
YOR027W	0.402	0.524	0.767175573	160.61	5.626	1.398.601.399	0.248	19	1.006	NO	6.1	589	39.66	stable	all alpha	enzyme regulator activity
YOR039W	0.324	0.714	0.453781513	71.47	3.891	0.946969697	0.15	29	0.737	NO	4.7	258	41.58	unstable	all alpha	enzyme regulator activity
YOR043W	0.53	0.5	1.06	95.49	2.56	0.755857899	0.154	8	1.007	NO	18.1	486	53.78	unstable	alpha beta	enzyme regulator activity
YOR046C	0.328	0.648	0.50617284	54.98	5.092	0.946073794	0.211	5	0.997	YES	0	482	32.38	stable	all alpha	helicase activity
YOR048C	0.435	0.573	0.759162304	16.52	2.422	2.364.066.194	0.164	2	1.002	YES	4.3	1006	50.33	unstable	alpha beta	hydrolase activity
YOR086C	0.516	0.584	0.883561644	40.78	3.092	1.246.882.793	0.22	2	1.028	NO	6.7	1186	28.94	stable	Membrane	molecular function unknown
YOR095C	0.308	0.7	0.44	54.44	3.711	1.506.024.096	0.248	1	0.996	YES	0	258	31.15	stable	alpha beta	isomerase activity
YOR116C	0.322	0.598	0.538461538	44.22	1.555	0.773993808	0.208	26	0.998	YES	1.4	1460	39.36	stable	all alpha	nucleotidyltransferase activity
YOR117W	0.322	0.303	1.062.706.271	91.82	4.944	0.786163522	0.195	22	0.996	YES	3	434	39.19	stable	alpha beta	peptidase activity
YOR142W	0.378	0.282	1.340.425.532	126.1	4.778	1.457.725.948	0.239	2	1.002	NO	7.3	329	28.67	stable	alpha beta	ligase activity
YOR151C	0.337	0.311	1.083.601.286	156.19	3.235	1.028.806.584	0.228	16	0.981	YES	0.9	1224	44.02	unstable	alpha beta	transferase activity
YOR155C	0.423	0.444	0.952702703	25.58	4.546	1.919.385.797	0.136	5	1.006	NO	2.7	450	56.98	unstable	alpha beta	hydrolase activity
YOR157C	0.279	0.754	0.370026525	46.34	4.793	0.77579519	0.179	9	0.999	YES	0	261	23.75	stable	alpha beta	hydrolase activity
YOR165W	0.452	0.565	0.8	104.34	3.501	0.444247001	0.182	0	1.012	NO	0	776	42.3	unstable	Membrane	molecular function unknown
YOR168W	0.391	0.395	0.989873418	320.35	3.929	102.145.046	0.269	0	0.983	YES	4.6	809	34.75	stable	alpha beta	ligase activity
YOR176W	0.376	0.72	0.522222222	80.72	7.469	0.778816199	0.169	9	1.001	YES	0	393	41.8	unstable	alpha beta	lyase activity
YOR187W	0.375	0.199	1.884.422.111	220.33	4.851	114.416.476	0.291	4	1.001	NO	4.3	1137	31.11	stable	alpha beta	translation regulator activity
YOR197W	0.427	0.456	0.936403509	61.07	4.216	0.930232558	0.172	9	1.01	NO	17.9	453	42.02	unstable	alpha beta	hydrolase activity
YOR201C	0.462	0.643	0.718506998	15.25	2.351	2.551.020.408	0.119	3	1.012	NO	2.7	412	38.21	stable	alpha beta	transferase activity
YOR204W	0.435	0.32	1.359.375	957.68	5.247	0.957854406	0.376	4	1.021	YES	23.2	604	42.63	unstable	all alpha	helicase activity
YOR207C	0.296	0.43	0.688372093	59.01	3.03	1.177.856.302	0.229	12	1.011	YES	1.5	1149	41.21	unstable	alpha beta	nucleotidyltransferase activity
YOR209C	0.375	0.671	0.558867362	73.55	5.121	0.883392226	0.207	2	1.006	NO	0	429	39.1	stable	alpha beta	transferase activity
YOR222W	0.3	0.757	0.396301189	33.82	4.265	0.895255148	0.181	1	0.995	NO	3.9	307	25.4	stable	alpha beta	transporter activity
YOR246C	0.501	0.518	0.967181467	64.63	1.692	0.448028674	0.12	1	1.001	NO	8.8	330	24.41	stable	Membrane	oxidoreductase activity
YOR253W	0.47	0.562	0.836298932	15.46	3.948	0.977517107	0.178	0	1.001	NO	0	176	38.05	stable	alpha beta	transferase activity
YOR259C	0.198	0.791	0.250316056	37.49	5.229	1.091.703.057	0.208	21	0.983	YES	5.7	437	40.23	unstable	all alpha	hydrolase activity
YOR260W	0.466	0.65	0.716923077	137.27	3.812	1.023.541.453	0.178	12	1.01	YES	13	578	44.03	unstable	alpha beta	enzyme regulator activity
YOR261C	0.3	0.743	0.403768506	26.91	3.223	3.344.481.605	0.194	26	0.984	YES	12.1	338	42.33	unstable	alpha beta	molecular function unknown
YOR272W	0.422	0.591	0.714043993	63.5	5.608	0.985221675	0.202	22	0.999	YES	2.4	460	38.23	stable	all beta	molecular function unknown
YOR283W	0.435	0.63	0.69047619	10.87	6.078	0.943396226	0.233	1	0.999	NO	0	230	37.42	stable	alpha beta	molecular function unknown
YOR303W	0.337	0.475	0.709473684	147.13	3.404	1.510.574.018	0.227	3	1	NO	0	411	31.98	stable	alpha beta	ligase activity
YOR323C	0.445	0.331	1.344.410.876	85.76	5.888	0.611620795	0.249	2	1.008	NO	0	456	28.41	stable	alpha beta	oxidoreductase activity
YOR326W	0.447	0.5	0.894	189.31	2.397	2.202.643.172	0.193	20	0.987	YES	9.4	1574	43.34	unstable	all alpha	motor activity
YOR335C	0.316	0.572	0.552447552	215.87	4.367	0.998003992	0.378	4	0.986	YES	4.5	958	30.45	stable	alpha beta	ligase activity
YOR341W	0.39	0.5	0.78	212.07	2.396	114.416.476	0.279	14	0.985	YES	4.1	1664	37.65	stable	Multidomain	nucleotidyltransferase activity
YOR361C	0.491	0.507	0.968441815	517.64	4.192	1.543.209.877	0.304	19	0.988	YES	2.5	763	37.75	stable	alpha beta	translation regulator activity
YOR370C	0.485	0.526	0.922053232	163.25	3.692	1.023.541.453	0.154	8	0.99	YES	5.1	603	36.3	stable	alpha beta	protein binding
YPL001W	0.454	0.556	0.816546763	7.34	4.894	0.888888889	0.155	9	0.756	NO	0	374	39.56	stable	alpha beta	protein binding

YPL012W	0.5	0.549	0.910746812	61.55	3.763	1.388.888.889	0.178	15	1.007	YES	4.6	1228	41.66	unstable	all alpha	molecular function unknown
YPL028W	0.43	0.224	1.919.642.857	701.45	5.258	1.177.856.302	0.366	5	1.006	YES	4	398	28.73	stable	alpha beta	transferase activity
YPL032C	0.57	0.634	0.899053628	40.53	1.82	0.537345513	0.164	5	1.011	NO	17.9	825	57.51	unstable	alpha beta	molecular function unknown
YPL043W	0.464	0.629	0.737678855	52.15	4.147	1.367.989.056	0.202	38	0.999	YES	12.1	685	43.68	unstable	alpha beta	RNA binding
YPL093W	0.342	0.664	0.515060241	535.93	4.567	0.953288847	0.36	27	0.979	YES	2.9	647	48.86	unstable	all alpha	Other
YPL106C	0.379	0.634	0.597791798	1938.74	5.682	0.846023689	0.521	4	0.954	NO	10.4	693	37.88	stable	alpha beta	Other
YPL111W	0.46	0.615	0.74796748	27.67	7.013	1.663.893.511	0.213	12	0.97	NO	0	333	28.74	stable	alpha beta	hydrolase activity
YPL117C	0.275	0.837	0.328554361	43.5	6.278	1.026.694.045	0.275	0	0.992	YES	6.6	288	33.76	stable	alpha beta	isomerase activity
YPL160W	0.388	0.451	0.860310421	22.66	4.336	1.126.126.126	0.3	4	0.988	YES	6.9	1090	42.63	unstable	alpha beta	ligase activity
YPL169C	0.529	0.615	0.860162602	57.25	4.022	1.579.778.831	0.12	11	0.986	YES	8.8	599	39.79	stable	alpha beta	structural molecule activity
YPL190C	0.537	0.625	0.8592	96.5	4.436	0.802568218	0.205	6	0.969	YES	35.2	802	73.41	unstable	alpha beta	RNA binding
YPL206C	0.54	0.58	0.931034483	39.06	5.335	0.867302689	0.148	0	0.989	NO	0	321	24.86	stable	alpha beta	hydrolase activity
YPL226W	0.41	0.419	0.978520286	147.73	3.242	1.094.091.904	0.304	2	0.995	NO	7.6	1196	37.67	stable	alpha beta	hydrolase activity
YPL235W	0.323	0.419	0.770883055	51.44	5.625	1.189.060.642	0.195	20	0.96	YES	4	471	38.8	stable	alpha beta	hydrolase activity
YPL237W	0.437	0.359	1.217.270.195	129.08	4.018	0.958772771	0.285	12	0.987	YES	12.3	285	43.36	unstable	alpha beta	translation regulator activity
YPL239W	0.394	0.661	0.596066566	10.15	5.729	0.738552437	0.152	0	0.983	NO	5	200	46.23	unstable	all alpha	molecular function unknown
YPR004C	0.426	0.545	0.781651376	41.65	3.299	1.324.503.311	0.146	1	0.995	NO	6.4	344	31.48	stable	alpha beta	molecular function unknown
YPR010C	0.369	0.312	1.182.692.308	213.44	3.255	0.908265213	0.223	22	0.983	YES	0	1203	38.8	stable	alpha beta	transferase activity
YPR016C	0.185	0.786	0.235368957	526.47	5.574	0.839630563	0.294	48	0.971	YES	0	245	36.46	stable	alpha beta	Other
YPR033C	0.338	0.58	0.582758621	291.49	4.969	0.975609756	0.295	4	0.714	YES	10.1	546	32.67	stable	alpha beta	ligase activity
YPR035W	0.241	0.645	0.373643411	151.08	7.633	1.322.751.323	0.525	1	0.984	YES	0	370	44.73	unstable	alpha beta	ligase activity
YPR037C	0.522	0.528	0.988636364	8.03	4.197	0.956937799	0.116	2	0.995	NO	0	196	45.55	unstable	Membrane	oxidoreductase activity
YPR041W	0.394	0.662	0.595166163	244.82	6.13	0.836120401	0.315	22	0.99	YES	7.2	405	40.05	unstable	alpha beta	enzyme regulator activity
YPR058W	0.354	0.686	0.516034985	39.63	3.044	0.883392226	0.145	3	0.985	NO	0	307	38.57	stable	alpha beta	transporter activity
YPR060C	0.445	0.475	0.936842105	6.83	5.55	1.081.081.081	0.217	0	0.964	NO	5.1	256	47.79	unstable	all alpha	isomerase activity
YPR088C	0.303	0.804	0.376865672	42.99	3.931	1.077.586.207	0.178	13	0.989	YES	18.7	541	48.42	unstable	all alpha	Other
YPR159W	0.434	0.373	1.163.538.874	182.43	4.785	1.340.482.574	0.189	8	0.99	NO	5.6	720	49.82	unstable	Membrane	hydrolase activity
YPR165W	0.334	0.141	2.368.794.326	316.99	4.293	0.874125874	0.267	26	1.01	YES	9.1	209	39.01	stable	alpha beta	hydrolase activity
YPR181C	0.42	0.4	1.05	342.87	4.126	1.193.317.422	0.229	18	0.813	YES	5.2	768	41.77	unstable	alpha beta	enzyme regulator activity
YPR191W	0.505	0.558	0.905017921	34.18	5.03	0.7390983	0.227	15	0.999	NO	0	368	28.56	stable	alpha beta	transporter activity

A.5 ARTIGO SUBMETIDO E ACEITE PARA PUBLICAÇÃO NO PROCEEDINGS OF MATHEMATICAL AND COMPUTATIONAL BIOLOGY – TORONTO, CANADA 2013

MINING THE CONSTRAINTS OF PROTEIN EVOLUTION[1]

FERNANDO ENCINAS

Laboratory of Computational and Systems Biology,
Oswaldo Cruz Institute, Rio de Janeiro, RJ, 21040-360, Manguinhos,Brazil

ANTONIO BASÍLIO DE MIRANDA[†]

Laboratory of Computational and Systems Biology,
Oswaldo Cruz Institute, Rio de Janeiro, RJ, 21040-360, Manguinhos,Brazil

The availability of different types of high-throughput data provides new opportunities for the identification of constraints that shape protein evolution; consequently, integrative computational approaches are essential to disclose the selective regimes that govern genomes. Combining text-mining analyses with other data mining techniques such as clustering and factor analysis, we have collected and analyzed data on various gene and protein characters to identify, classify and reveal existing associations between characters that may favor or hinder the rate at which proteins evolve. The use of latent constructs as an integrative procedure aimed to explain from a system perspective the relationships and the strength of these genome-wide characters allowed us to find that, at least for our data set, expression and structural constructs synthesize more the information of our data set in comparison to functional constructs. Samples from a posterior distribution of a Bayesian model showed that, at the level of an effective and accurate protein translation system, synonymous substitutions and translational efficiency are correlated and both influence the system positively whereas the structure instability and the dispensability of a protein have, yet small, a negative influence on it. Overall, this work presents an integrative methodology intended to make the most of the available genomic data and describes an alternative framework to size the strength and links between determinants of protein evolution.

Introduction

The causes of variation in protein evolutionary rates have been a recurring topic of interest in the field of evolutionary biology [1,2,3]. Various comparative genomic analyses allowed the identification of individual factors, functional and structural, that favor or hinder the rate at which substitutions accumulate at nucleotide level [4,5,6]. Among these, although some examples against exist [7], gene expression has been indicated as major determinant of protein evolution [8,9].

The access to different types of biological information confirmed the complexity of organisms as living systems [10] and blurred the phenotypic boundaries at which selection operates [11]. Therefore, in the light of the ever-growing amount of high-throughput experimental data, there is a need to review the constraints that govern evolutionary change and to integrate related data to tackle protein evolution from an integrated perspective.

Integration of related data is particularly fruitful as it brings out the real value of individual data sets; however, to make this integration feasible and meaningful, it is necessary the application of advanced computational methods accompanied by mathematical and statistical approaches adequately braced with a theoretical framework [12].

[*] This work was supported by the Institutional Cooperation of the Institute Oswaldo Cruz
[†] Corresponding author: antonio@fiocruz.br

Data mining as an applied science is a computer assisted process of analyzing large amounts of data to summarize it into valuable information [13]. Through a wide range of techniques, data mining approaches allow the recognition of patterns that are not instantly apparent and have the flexibility to offer both individual and system-level explanations [14].

In this work we present a combined methodology that, starting with a text-mining analysis, collected data on genome-wide variables that may constitute determinants of protein evolution. Hierarchical clustering and advanced factor analyses were used to explain the structure of the data set at a higher level and finally, a Bayesian factor model was tested to estimate what would be the components of an effective-accurate protein translation system.

Methods

2.1. *Text Mining*

Sixty one PDF research articles on protein evolution were manually downloaded from PUBMED and converted to plain text. An in-house code implemented in C language was used to process these plain texts by extracting sections of interest such as abstract, introduction and discussion. Resulting text files formed the document collection that was analyzed by the tm package [15] in R environment [16]. Text transformation, corpus construction and association between frequent terms were used to process the information form texts.

2.2. *Data collection*

We collected expression information including mRNA levels, translational efficiency and protein abundance for genes encoded in the genome of *Saccharomyces cerevisiae* for which comparative transcriptome/proteome analyses were conducted in [17]. Functional data consisting of dispensability and number of interactions were downloaded from (http://chemogenomics.stanford.edu/supplements/01yfh/files/orfgenedata.txt) and Database of Interacting Proteins (http://dip.doe-mbi.ucla.edu/dip/) respectively. Structure-related information consisting of native structure classification, low complexity percentage and protein length were retrieved from Pedant Database (http://pedant.helmholtz-muenchen.de/genomes.jsp?category=fungal). Finally, all genes were classified according to Gene Ontology classification using the Slim Mapper of *Saccharomyces* Genome Database (SGD) (http://www.yeastgenome.org/).

Pairs of orthologous genes between *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* were found using the stand-alone version of the InParanoid algorithm [18] and aligned using the ClustalW 2.0 [19] program with default parameters. Evolutionary rates, number of non-synonymous substitutions per synonymous site (dN) and synonymous substitutions per synonymous site (dS), between each ortholog pair, were estimated using the method of Nei and Gojobori implemented in MEGA 4 [20].

2.3. *Data Mining*

Summarization can be viewed as a compression of data into a smaller set of patterns retaining the maximum informative representation. We have used the following data mining techniques to summarize our data set:

2.3.1. *Hierarchical clustering of variables*

An ascendant hierarchical algorithm was used to arrange qualitative and quantitative variables in clusters of decreasing homogeneity. The homogeneity of a cluster is defined as the sum of correlation ratios (for qualitative variables) and the squared correlation (for quantitative ones) to a synthetic variable. The R package ClustOfVar [21] was used to implement the algorithm.

2.3.2. *Multiple Factor Analysis*

Multiple Factor Analysis (MFA) makes the synthesis of weighted Principal Component Analysis for quantitative variables and weighted Multiple Correspondence Analysis for qualitative variables making possible the analysis of variables structured into groups of related nature. Functions from the FactoMineR package [22] were used to perform MFA in six groups of variables arranged according to Table 1.

2.3.3. *Bayesian Factor Analysis*

Having a certain set of observed variables, Bayesian Factor Analysis incorporates a prior to construct a measurement model that estimates the indeces of a latent construct. Markov Chain Monte Carlo algorithms are used to fit the factor model sampling the factor loadings from the posterior distribution. The main idea is to explain the relationships between a set of observed variables in terms of an unobserved variable via a relatively parsimonious model. Software for fitting the model is available in the MCMCpack [23] package for R and detailed derivation of factor analysis model and posterior inference can be found in [24].

Bayesian perspective depends on a prior however, we did not constrain the elements to the factor, the prior mean of each element and prior precision were assumed to be 0. Initial 1000 MCMC scans were discarded as burn-in and storing every 100th scan, 100000 iterations were necessary for the Markov Chain to converge. Heidelberg and Welch's convergence test was used to verify if the sample values come from a stationary distribution.

3. Results

3.1. *Genomic variables derived from text identifiers*

An essential task, even for the simplest text mining analysis, is finding the terms that recur in a collection of documents. This enables the condensation of the whole content of information into a limited number of words. Frequent terms represent the identifiers of a collection therefore, finding significant associations between them (i.e., terms which co-occur) makes it possible to group and organize concepts to another level of valuable information.

We have combined term frequency and term association analysis in a set of research articles to find new, potential constraints of protein evolution. Thirty-one most frequent terms condensed the information of the texts and some of them visibly implied certain genomic information (Appendix A.1).

In terms of co-occurrence counts, some terms presented significant correlations (Fig.1) that were very useful to support the intuitive attribution of one or more of them to a specific gene or protein character.

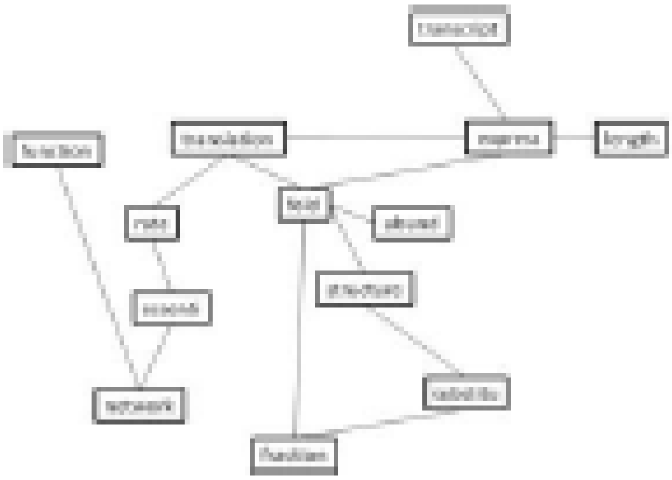


Figure 1 Term association network. Edges indicate significant co-occurrences between terms and provide the framework for concept classifications and collection identification.

As a result, thirteen genomic variables, among gene and protein characters, were identified as prospective constraints of protein evolution and included as the focus of study in subsequent analyses. Table 1 presents the terms, the data type, nature and brief description of the genomic variables considered in the study.

3.2. *Pair-wise analyses reveal existing relationships between various genomic variables*

We collected or calculated the values of genomic variables listed in Table 1 for 442 protein-coding genes in the genome of the model organism *Saccharomyces cerevisiae* as detailed in Methods section.

We were especially interested in analyzing the behavior of "new" characters that might relate either to evolutionary variables or to expression variables.

Table 1 Detailed description of the origin, type and nature of genomic information

Stem term	Gene/protein feature	Variable type	Nature
rate, substitu	number of synonymous substitutions (dS)	continuous	Evolutive
rate, substitu	number of non synonymous substitutions (dN)	continuous	Evolutive
express	mRNA level	continuous	Expression
abund	protein level	continuous	Expression
translation	translation efficiency	continuous	Expression
length	protein length	continuous	Structural
structure	native structure	categorical	Structural
structure	instability index	continuous	Structural
structure	Stability	categorical	Structural
network	number of interactions	continuous	Functional
region, structure	low complexity percentage	continuous	Structural
essenti	Essentiality	categorical	Functional
essenti	Dispensability	continuous	Functional

We were especially interested in analyzing the behavior of genomic characters that might relate either to evolutionary variables or to expression variables. Thus, as shown in Figure 2**,** a strong negative correlation (- 0.3307, *p*< 9.55e-13) is evident between the level of expression (mRNA level) and the number of non-synonymous substitutions (dN) (Fig.2A) and between translation efficiency and dn (-0.2467, p<1.48e-07) (Fig. 2B).

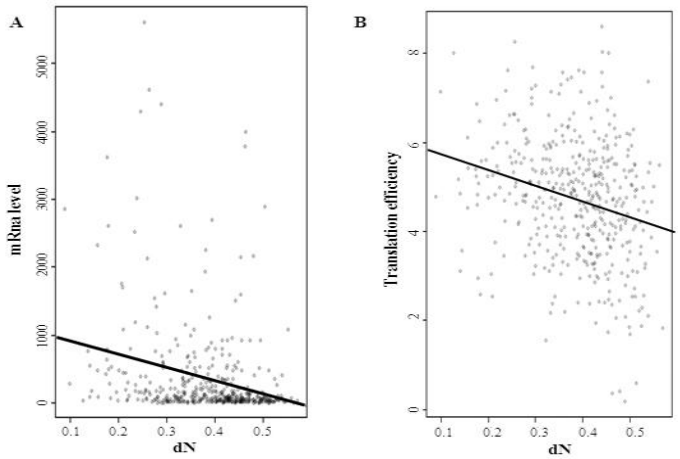


Figure 2 Negative correlations between mRNA level and dN (A) and translation efficiency and dN (B)

Turning into some of the "new" genomic characters we found in text-mining analysis, the instability index of a protein, a structure-related variable, presented high positive correlation with dN and a strong negative correlation with some expression variables such as translation efficiency (Appendix B.1).

Although these preliminary results demonstrate the potential of text-mining approaches to generate novel information and reinforce the notion that more and strong genomic constraints do exist, they poorly contribute to our understanding on the evolution of proteins from an integrated perspective.

3.3. *Clustering of variables reveals the underlying structure of the data*

As clustering genomic variables in homogeneous groups would provide meaningful global information, we applied a hierarchical clustering algorithm based on agglomerative schemes to the mixture of quantitative and qualitative variables from our data set.

Aggregation levels demonstrated that four clusters would be enough to reveal the structure of the data (Appendix C.1) thus, as depicted in the dendrogram of Fig. 3, most variables appeared to form clusters easily defined by the nature of the correlating variables.

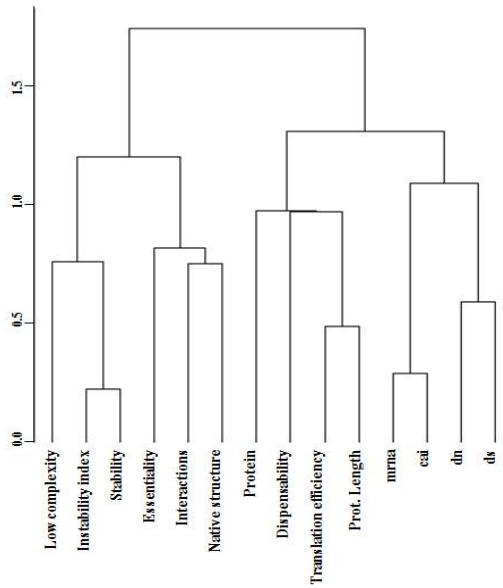


Figure 3 Hierarchical clustering of variables. Four groups of correlating variables reveal the structure of the data set

In terms of homogeneity, low-complexity, instability index and stability, three structure-related variables, clearly grouped in the same cluster. Essentialiy and number of interactions grouped together with native structure in a second cluster. Protein abundance, translation efficiency and protein length, all related to the translation machinery linked in a third cluster. Finally, evolutive variables dS and dN grouped together with a expression related variable, mRNA level. Individual squared loadings for each cluster can be found in Appendix C.2.

3.4. *Latent constructs are useful to integrate genomic data and provide a descriptive system perspective*

Grouping genomic variables into clusters allowed us to grasp the underlying structure of our data set; nevertheless, no information is provided about the type or direction (positive or negative) of existing relationships between variables.

Aimed to analyze simultaneously multiple sets of variables, Multiple Factor Analyses (MFA) use an arrangement of variables in groups of related nature to evaluate the influence of each group and to reveal if there is any relationship between such groups. A descriptive concept or latent construct can be associated to each group in order to attain a system-level interpretation.

Six groups of related genomic variables were created as detailed in Methods section and Table 1 to be analyzed by functions included in the package FactoMiner [22]. Figure 4 shows the quality representation of each group of variables clearly separated in the axes projection.

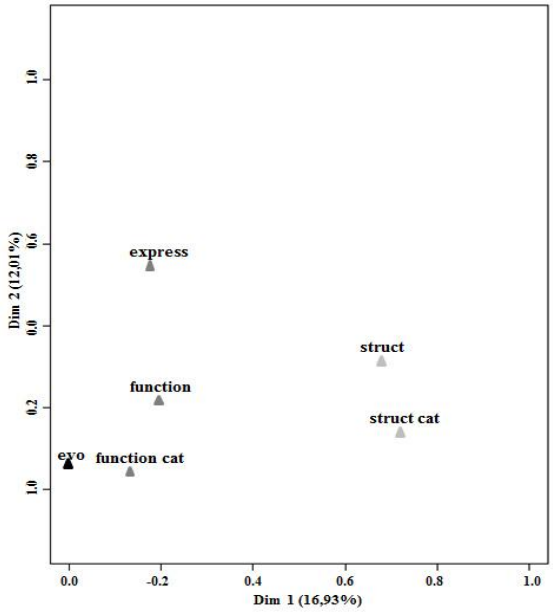


Figure 4 Quality representation of latent constructs. Related data on three major determinants of protein evolution can be integrated using latent constructs that synthesize distinct information reliably.

The distance between groups suggests, as we expected, that each of them represents distinctive but integrated information on three major determinants of protein evolution: structure, expression and function. Structural constructs (struct and structcat) appeared to have high coordinates on the first axis, whereas expression construct (express) had the highest coordinates to the second axis. Both located distant from the evolution construct (evo), which has been set as supplementary group, and from the point of origin showing that these groups of variables helped the most in the synthesis of the information. Function constructs (function and functioncat) on the other hand, although separated equally, they presented low coordinates on the first two axes, consequently little power of discrimination.

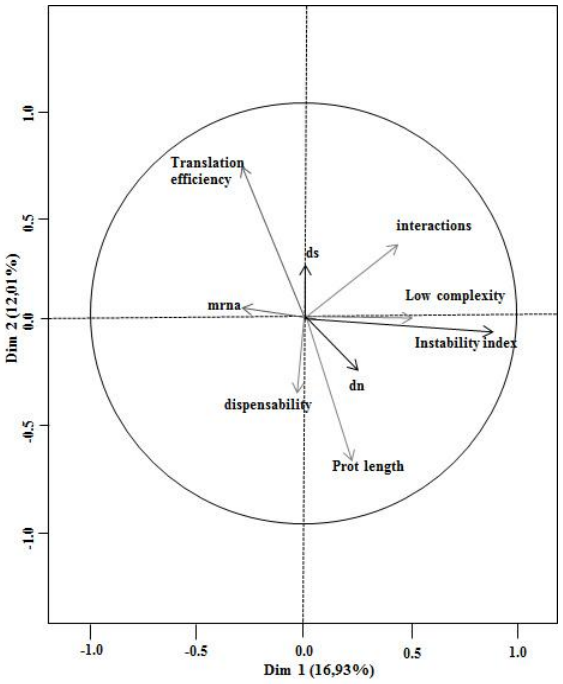


Figure 5 Circle of correlations. The individual coordinates show graphically the relationships between variables

Individual coordinates for members of each group provide the definitive descriptive system perspective proposed throughout the work. Figure 5 presents a plot of the factorial map of a correlation circle in which it is noticeable, on one hand, the opposition between expression variables and the number of non-synonymous substitutions; on the other hand, the high correlation between structure-related variables (low complexity percentage and instability index) and finally, the positive association between translation efficiency and synonymous substitutions both opposing to the length of a protein and to its dispensability. Information on Eigenvalues and cumulative percentage of variance can be found in the Appendix D.1.

3.5. *Model estimates show positive and negative contributors to an effective-accurate protein translation system*

To study the intricate relationships at the level of a particular system, we used a Bayesian Factor Analysis that, by using a prior and a given set of variables, it allows the construction of measurement models to estimate the indices of a latent construct. Markov Chain Monte Carlo algorithms are used to sample the factor loadings from a posterior distribution.

We used five genomic variables (number of synonymous substitutions, translational efficiency, protein abundance, dispensability and instability index) to construct the indices of a latent construct intuitively identified with an effective-accurate protein translation system. The goal of the current model is to capture patterns of association between the variables and the latent construct.

In principle, a Bayesian perspective depends on a prior, however we did not constrain any of the variables to identify the model. 100000 iterations were enough to reach stationarity as verified by diagnostic analysis (Methods)

Table 2 presents a summary of the posterior distribution of factor loadings and psi-uniqueness as part of the model's output. In line with our expectations, the factor loading of translational efficiency resulted high indicating a strong association between the efficiency at which a protein is translated and the latent construct. In the same line, although showing a relatively lower factor loading, the number of synonymous substitutions indicated a positive influence to the latent construct as well.

Table 2 Posterior distribution of factor loadings and uniqueness of the Bayesian factor analysis

	Factor loading	Psi-uniqueness
synonymous substitutions	0.4121	0.6921
instability index	-0.2134	0.9548
translation efficiency	0.8783	0.2129
protein level	-0.1410	0.9826
Dispensability	-0.0995	0.9954

In general, factor loadings tend to increase as more iterations are specified in the MCMC; consequently, in terms of type of association, the sign of a factor loading provided the information on the influence of each variable to the latent construct. As showed in Table 2, instability index, protein level and the dispensability of a protein were all estimated to be negative contributing negatively to the translation system.

4. Discussion

Research articles constitute the primary source of biological information. For years, scientific literature repositories have accumulated information on studies interested in the interplay between genotype and phenotype that identified and correlated individual genomic attributes that determine selective constraints. Consequently, as the rate of textual information grows, new computational methods are required to discover hidden, unsuspected and potentially valuable information.

Text mining has emerged as a leading-edge technology that takes advantage of techniques of information retrieval, natural language processing and data mining, to cope with the non-trivial task of

dealing with the ambiguity in language and the unstructured nature of written documents [25]. In biology, its applications vary from drug discovery [26] and disease-gene associations [27] to the systematic review of protocols and analysis of trends in molecular biology [28].

As pointed previously, the most elementary task in text analysis is to extract the terms that recur in a collection of documents. However, in practice, low frequency terms occur in few documents whereas highly frequent terms tend to pollute the selection of key identifiers. Therefore, the number of text included in a collection, the transformation of documents, the removal of contaminant terms and the overall pre-processing in text mining analysis constitute crucial steps to obtain satisfactory results.

Assigning the identifiers of our text collection to gene or protein features, we have been able to distinguish variables that, in the light of pair-wise correlation analysis, appear to be unacknowledged constraints of protein evolution. The instability index, the translation efficiency and percentage of low complexity regions in a protein strongly correlate with the number of non- synonymous substitutions (dN) accumulated.

In the same direction, our results showed that the level of activation of a gene, expressed by its mRNA level, also correlate negatively with dN, supporting the view that highly expressed genes tend to evolve at a slow rate. It has been suggested that evolution progresses through changes in protein expression rather than sequence [29]; therefore, gene expression constitutes the "key" element in our understanding of protein evolution.

While this "key" is generally interpreted as the unequivocal negative association between these variables (dN and mRNA level), it can be also argued that it holds a simplistic view of what gene expression really represents and especially that restricts the action of selection to a narrow margin. Gene expression can be explained by the level at which one exon is transcribed, by the number of translations per transcript or by the level of structurally functional proteins in the cell. Thus, transcription, translation and protein abundance might be important to different extents and selection may have a role at different stages accordingly [30].

Due to the requirement to form and maintain the definitive active (as in the case of enzymes) site that probably exerts a strong selective pressure on a protein to adopt just one stable and conserved fold, protein structures are generally regarded as "fossil records" of molecular evolution [31]. However, as more protein structures become available and more structural genomics projects are generating new and unprecedented information, a major biological question is how a system's physical properties influence its capacity to evolve.

On the one hand, it has been shown that contrary to the traditional view that protein function equates with a stable three-dimensional structure, many gene sequences in eukaryotic genomes encode large segments or even entire proteins that lack a well-structured three-dimensional fold and moreover, some of these regions can be highly conserved between species [32, 33]. On the other hand, there is strong evidence that the capacity of one protein to evolve is enhanced by the mutational robustness conferred by extra stability [34].

As we see, the availability of different types of high-throughput biological data serves as evidence of the complexity that living organisms have reached in millions of years under the influence of selective forces that shaped their evolutionary history. However, the real informative value of individual data sets is truly appreciated only if these are combined or integrated in a single framework.

Data-mining techniques can provide such a framework and constitute an ideal option for the analysis of "different-but related" data sets. Unfortunately, most traditional algorithms in data mining are limited to handling datasets that contain either continuous or categorical variables, reducing thus the choices of researchers to discard or to discretize some of them and making it impossible to uncover the multidimensional structure of the observed data. Our work, as it happens in most of real life examples, is composed by a mixture of continuous and categorical attributes; therefore, to fully exploit the characteristics of the entire data set, we relied heavily on methods that are appropriated to deal with mixed types of attributes.

Initially intended to serve as a simple exploratory or pre-processing step, the hierarchical clustering of variables resulted especially useful to reveal the intrinsic structure of our data set. We have been able to recognize clusters of genes' or proteins' features that made recognizable not only the nature and the information that grouping variables bring, but also the associations that may exist between them. While the identification of clusters related to structural information and evolutive nature was straightforward, the cluster formed by variables "dispensability", "translation efficiency", "protein abundance" and "protein length" do not share any obvious nature for grouping and suggest the existence of an orchestrated interplay of diverse components whose recognition would greatly facilitate the understanding of a biological system as a whole.

Latent concepts play important roles in the theoretical work of many fields [24,35] and we took advantage of their virtue to act as components of both individual and system-level explanations to review the classic views of protein evolution in the light of the genomic data available.

The classic view would state that protein evolution is basically affected by selection acting on protein structure and function; moreover, mRNA level, as proxy for gene expression, has been pointed to have a major influence on the evolution of the corresponding gene. In contrast, our approach prioritized the quest for general over particular determinants of protein evolution.

A key process in the biology of a cell is the synthesis of proteins with high efficiency and fidelity. Thus, in recent years, we have witnessed an increased interest to understand the evolutionary mechanisms that led to the adaptation of the protein translation system [36,37].

The study of complex systems begins with the identification and simplified description of the individual components of such a system. We used a Bayesian Factor Analysis to identify the components of what would be an efficient and accurate (adapted) protein translation system and found that, according to our model, synonymous substitutions and translation efficiency constitute positive contributors to an adapted translation system, whereas dispensability, instability index and the abundance of a protein negatively associate with the system.

Although synonymous substitutions have been traditionally regarded as samples of neutral evolution, in last years, studies have shown they exert a profound effect in the efficiency of the translation system since certain codons are translated faster or more accurately than others [38]. Synonymous codons also appear to have different influences on the co-translational folding process of nascent proteins [39].

Recently, a study from Stevens *et al.* (2013) [40] estimated the translation efficiency for a set of genes in different cell lines combining information from mRNA levels and protein stability, supporting, to a certain extent, the inclusion of the instability index to the construction of our translation model.

Considering the importance for an organism to faithfully count with functional proteins, the unexpected negative association between protein abundance and an efficient-accurate translation system initially suggests that a more precise model specification should constrain this variable parameter to load positively on the factor. This negative association however, can also be explained by the delaying effect that the kinetic translational control exerts through clusters of rare codons that ultimately favors fidelity over efficiency.

5. Conclusions

Life sciences are facing the challenge of handling and analyzing biological information through the use of more innovative computational methods to respond the growing need of making sense of large amounts of experimental data. Integration of related data is useful to this purpose as it brings out the real value of individual data sets and, if linked to a theoretical framework, it provides the system-level perspective to review classical assumptions and test new hypothesis.

In this work, combining text-mining techniques with simple correlation analyses we have been able to identify genomic features that appear to be overlooked when studying the rates of protein evolution and the targets of selective forces. Translation efficiency, structural instability and low complexity regions showed strong correlation with the rate at which a protein evolves.

Latent constructs were used as an alternative to integrate related genomic information and to approach the evolution of biological organisms as systems formed by different components. We could recognize clearly distinct constructs that each in turn bring different information and found that, in general, expression and structural constructs explain more our data set in comparison to functional constructs. Overall, our results suggest that rather than taking mRNA levels as major determinants of protein evolution, other expression related should be considered.

A Bayesian factor model allowed us to identify the estimates of a latent construct interpreted as an effective and accurate translation system and, although our model may lack the theoretical rigor, in particular, it helped us to grasp global patterns of the system, the positive association of synonymous substitutions and translational efficiency with the construct and finally, in general, it demonstrates the applicability of similar approaches for the analysis of protein evolution.

6. Appendix

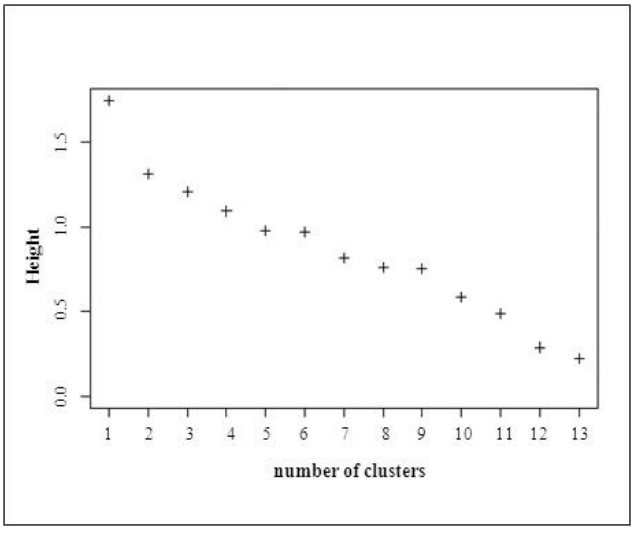
A.1. List of most frequent terms in the collection of documents

```
 [1] "chang"        "correl"       "data"
 [4] "differ"       "effect"       "evolut"
 [7] "evolutionari" "evolv"        "express"
[10] "figur"        "function"     "gene"
[13] "genom"        "interact"     "level"
[16] "mutat"        "network"      "ortholog"
[19] "protein"      "rate"         "relat"
[22] "residu"       "result"       "select"
[25] "sequenc"      "site"         "speci"
[28] "structur"     "studi"        "use"
[31] "yeast"
```

B.1. Matrix of correlations

Variable	ORF length	dN	dS	dN/dS	mRNA level	Translational efficiency	Protein abundance	CAI	Number of interactions	Dispensability	% Low complexity	Protein length	Instability index
ORF length	100	17.74	-19.12	14.13	3.17	-51.78	6.31	-12.03	3.69	8.25	2.54	99.35	8.33
dN		100	-41.15	62.6	-26.68	-25.06	3.49	-49.6	-6.77	1.42	9.16	17.54	22.22
dS			100	-89.4	-5.13	10.65	2.32	-1.86	-3.11	-3.27	1.54	-20.14	3.17
dN/dS				100	-9.77	-13.4	-0.74	-19.56	0.37	2.32	2.16	15.35	4.91
mRNA level					100	8.15	-2.84	71.14	-3.53	-2.47	3.73	3.25	-17.54
Translational efficiency						100	-11.63	29.92	0.1	-9.57	-10.38	-51.33	-18.18
Protein abundance							100	-7.38	-2.01	-2.06	1.64	6.32	9.66
CAI								100	-5.2	-4.14	5.78	-11.67	-27.04
Number of interactions									100	-1.25	9.52	3.7	9.93
Dispensability										100	-4.44	8.58	-5.06
% Low complexity											100	2.66	38.93
Protein length												100	7.89
Instability index													100

C.1. Aggregation levels for number of clusters of variables



C.2. Squared loadings corresponding to four clusters of variables

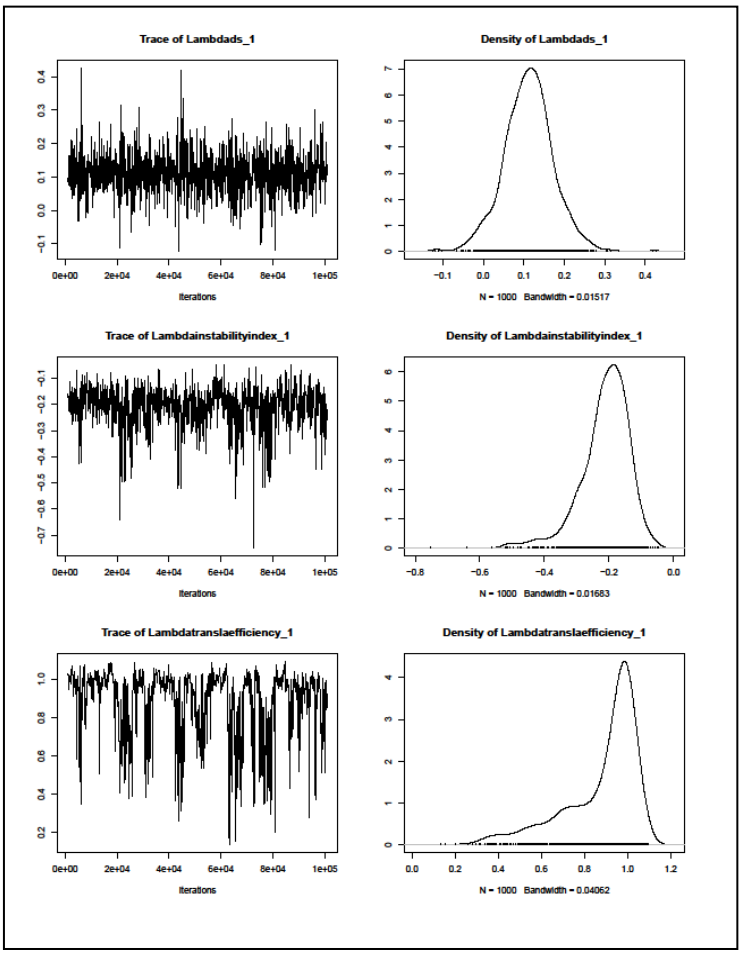
Cluster 1	Squared loading	Cluster 2	Squared loading
dN	0.53571258	Translation efficiency	0.72967024
dS	0.05580212	Protein level	0.06677992
mRNA	0.63455515	Dispensability	0.06813235
CAI	0.80514858	Protein length	0.70412936

Cluster 3	Squared loading	Cluster 4	Squared loading
Number of interactions	0.5174661	Low complexity	0.3843773
Essentiality	0.4435748	Instability index	0.8422532
Native structure	0.4694705	Stability	0.7914906

E.1. Heidelberg and Welch's convergence test for Bayesian Factor Model

	Stationary	Iteration	p-value
Lambda-dS	Passed	1	0.243
Lambda instability índex	Passed	1	0.180
Lambda- translation efficiency	Passed	1	0.122
Lambda-protein level	Passed	1	0.165
Lambda-dispensability	passed	1	0.584
Psi-dS	Passed	1	0.608
Psi-instability index	Passed	1	0.219
Psi-translation efficiency	passed	1	0.104
Psi-protein level	Passed	1	0.380
Psi-dispensability	Passed	1	0.454

E.2. Posterior Densities for some variables in the Bayesian Factor Model



7. References

1. C. Pál, B. Papp, MJ.Lercher, *Nat. Rev. Genet.* 7, 5 (2006)
2. JI. Lucas-Lledó, M. Lynch, *Mol. Biol. Evol.* 26, 5 (2009)
3. X. Du X, DJ. Lipman, JL. Cherry, *Genome Biol. Evol.* 5, 3 (2013)
4. S. Vieira-Silva, M. Touchon, SS. Abby, EP. Rocha, *Proc. Natl. Acad. Sci.* 108,50 (2011)
5. J. Coulombe-Huntington, Y. Xia, *PLoS Comput. Biol.* 8, 10 (2012)

6. S. Chakraborty, B. Kahali, TC. Ghosh, *BMC Syst. Biol.* 12;4 (2010)
7. I. Tirosh, N. Barkai, *Trends Genet.* 24, 3 (2008)
8. DA. Drummond, A. Raval, CO. Wilke, *Mol. Biol. Evol.* 23, 2 (2006)
9. JF. Gout, D. Kahn, L. Duret, *PLoS Genet.* 6, 5 (2010)
10. B. Berger, J. Peng, M. Singh, *Nat. Rev. Genet.* 14, 5 (2013)
11. E. Koonin, Y. Wolf, *Nat. Rev. Genet.* 11, 7 (2010)
12. PV. Gopalacharyulu, E. Lindfors, C. Bounsaythip, T. Kivioja, L. Yetukuri, J. Hollmén, M. Oresic, *Bionformatics* 21, 1 (2005)
13. H. Bensmail, A. Haoudi, *J Biomed. Biotech.* 2, (2005)
14. D. Rebholz-Schuhmann, A. Oellrich, R. Hoehndorf, *Nat. Rev. Genet.* 13, 12 (2012)
15. I. Feinerer, K. Hornik, D. Meyer, *J. Stat. Soft.* 25, 5 (2008)
16. R. Ihaka, R. Gentleman, *J. Comp. Graph. Stat.* 5, 3 (1996)
17. VL. MacKay, X. Li, MR. Flory, E. Turcott, GL. Law, KA. Serikawa, XL. Xu, H. Lee, DR. Goodlett, R. Aebersold, LP. Zhao, DR. Morris, Moll. Cell. *Proteomics.* 3, 5 (2004)
18. G. Ostlund, T. Schmitt, K. Forslund, T. Kostler, DN. Messina, S. Roppa, O. Frings, EL. Sonnhammer. *Nucleic Acids Res.* 38 (2010)
19. JD. Thompson, DG. Higgins, TJ. Gibson, *Nucleic Acids Res.* 22 (1994)
20. K. Tamura, J. Dudley, M. Nei, S. Kumar, *Mol. Biol. Evol.* 24, 8 (2007)
21. M. Chavent, V. Kuentz-Simonet, B. Liquet, J. Saracco, *J. Stat. Soft.* 50, 13 (2012)
22. S. Le, J. Josse, F. Husson, *J. Stat. Soft.* 25, 1 (2008)
23. D. Martin, M. Quinn, Jong Hee Park, *J. Stat. Soft.* 42, 9 (2011)
24. M. Quinn, *Pol. Anal.* 12 (2004)
25. D. McDonald, U. Kelly, *JISC* (2012)
26. C. Plake, M. Schroeder, *Curr. Pharm. Biotechnol.* 12, 3 (2011)
27. H. Al-Mubaid, RK. Singh, *Int. J. Bioinform. Res. Appl.* 6, 3 (2010)
28. M. Krallinger, RA. Erhardt, A. Valencia, *DDT.* 10, 6 (2005)
29. C. Bustamante, A. Fledel-Alon, S. Williamson, R. Nielsen, MT. Hubisz, S. Glanowski, DM. Tanenbaum, TJ. White, JJ. . Sninsky, RD. Hernandez, D. Civello, MD. Adams, M. Cargill, AG. Clark, *Nature.* 437, 7062 (2005)
30. EP. Rocha, *Trends Genet.* 22, 8 (2006)
31. A. Andreeva, AG. Murzin, *Curr. Opin. Struct. Biol.* 16, 3 (2006)
32. J. Nilsson, M. Grahn, AP. Wright, *Genome Biol.* 12, 7 (2011)
33. HJ. Dyson, PE. Wright, *Nat. Rev. Mol. Cell. Biol.* 6, 3 (2005)
34. JD. Bloom, ST. Labthavikul, CR. Otey, FH. Arnold, *Proc. Natl. Acad. Sci.* 103, 15 (2006)
35. K. Bollen, Annu. *Rev. Psychol.* 53, 605 (2002)
36. D. Herman, CM. Thomas, DJ. Stekel, *PloS ONE.* 7, 11 (2012)
37. M. Gilchrist, P. Shah, R. Zaretzki, *Genetics.* 183 (2009)
38. S. Shabalina, N. Spiridonov, A. Kashina, *Nucleic Acids Res.* 41, 4 (2013)
39. G. Zhang, M. Hubalewska, Z. Ignatova, *Nature Struct. Mol. Biol.* 16, 3 (2009)
40. S. Stevens, C. Brown, *PloS ONE.* 8, 2 (2013)