OTÁVIO JOSÉ BERNARDES BRUSTOLINI

PREDIÇÃO *IN SILICO* DE PROTEÍNAS EXTRACELULARES DE *Kluyveromyces lactis* E SUAS RELAÇÕES COM FATORES TRANSCRICIONAIS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Microbiologia Agrícola, para a obtenção do título de Magister Scientiae.

VIÇOSA
MINAS GERAIS - BRASIL
2008

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T

B912p
2008

Brustolini, Otávio José Bernardes, 1978-
Predição *in silico* de proteínas extracelulares de *Kluyveromyces lactis* e suas relações com fatores transcricionais / Otávio José Bernardes Brustolini. – Viçosa, MG, 2008.
v, 40f. : il. (algumas color.) ; 29 cm.

Texto em português e inglês.
Orientador: Flávia Maria Lopes Passos.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. *Kluyveromyces lactis*. 2. Leveduras (Fungos).
3. Proteínas microbianas. 4. Enzimas microbianas.
I. Universidade Federal de Viçosa. Departamento de Bioquímica e Biologia Molecular. Programa de Pós-Graduação em Microbiologia Agrícola. II. Título.

CDD 22. ed. 579.76

OTÁVIO JOSÉ BERNARDES BRUSTOLINI

PREDIÇÃO *IN SILICO* DE PROTEÍNAS EXTRACELULARES DE *Kluyveromyces lactis* E SUAS RELAÇÕES COM FATORES TRANSCRICIONAIS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Microbiologia Agrícola, para a obtenção do título de *Magister Scientiae*

APROVADA: 31 de julho de 2008

_____	_____
Hilário Cuquetto Mantovani	Karla Suemy Clemente Yotoko
_____	_____
Cosme Damião Cruz (Coorientador)	Luciano Gomes Fietto (Coorientador)

_____
Flávia Maria Lopes Passos
(Orientadora)

SUMÁRIO

RESUMO

BRUSTOLINI, Otávio José Bernardes, M.Sc., Universidade Federal de Viçosa, julho de 2008. **Predição *in silico* de Proteínas Extracelulares de *Kluyveromyces lactis* e suas relações com fatores transcricionais.** Orientadora: Flávia Maria Lopes Passos. Co-orientadores: Cosme Damião Cruz e Luciano Gomes Fietto.

O banco de dados da *Kluyveromyces lactis* constituído de 5327 seqüências de proteínas (http://cbi.labri.fr/Genolevures) foi submetido a quatro algoritmos de predição para identificar o potencial secretome extracelular. O primeiro,*SignalP* v3 (http://www.cbs.dtu.dk/services/SignalP-3.0), que identifica a presença de peptideo sinal na porção N-terminal e o sítio de clivagem da peptidase sinalagrupou 698 proteínas.Deste grupo, o *Phobius*(http://phobius.sbc.su.se), que prevê a topologia de domínios transmembranas a partir das sequencias primárias, indicou 260 sem domínios transmembranas.Outros dois algoritmos, *big-PI predictor*(http://mendel.imp.ac.at/gpi/gpi_server.html),capaz de reconhecer marcas de ancoras GPI (Glicosilfosfatidilinositol) e *WoLF PSORT*(http://www.genscript.com/psort/wolf_psort.html)capaz de identificar assinaturas para a localização em compartimentos subcelulares apontaram 236 proteínas sem ancoras GPI e 101 endereçadas ao meio extracelular. Como controle positivo, os mesmos algoritmos foram testados e predisseram corretamente 95 proteínas de leveduras Saccharomycetes encontradas nos bancos de dados públicos (NCBI e UNIProt) e anotadas como extracelulares. Como controle negativo foram preditas como intracelular 95 seqüências aleatórias do banco de dados da *K. lactis*. O grupo controle positivo e o grupo predito foram comparados pelo teste estatístico T^2 de Hotelling. Não foram evidenciadas diferenças significativas entre os valores das médias dos grupos.A condição fisiológicana qual estas proteínas extracelulares são expressas foi analisada relacionando suas seqüências promotoras com os

fatores transcricionais ortólogos da *Saccharomyces cerevisiae*. A metodologia aplicada foi o "Yeastract" (http://www.yeastract.com) que localiza sítios de ligação ao DNA dos fatores transcricionais de *S. cerevisiae*nas seqüências promotoras dos ORFs das proteínas preditas como extracelulares. A condição fisiológica que favorece a expressão para o meio extracelular foi obtida pela pesquisa dos termos descritos pelo "Gene Ontology"(http://www.geneontology.org). Os fatores transcricionais que mais se relacionam com as seqüências preditas foram aqueles associados com resposta a estresse. Também foi indicado que o estresse ácido e limitação de nitrogênio (aminoácidos) exercem influência na expressão das proteínas extracelulares.

ABSTRACT

BRUSTOLINI, Otávio José Bernardes, M.Sc., Universidade Federal de Viçosa, July 2008. **Computational analysis of the interaction between the transcriptional factors and the predicted secreted proteome of the yeast *Kluyveromyces lactis*.** Adviser: Flávia Maria Lopes Passos. Co-Advisers: Cosme Damião Cruz and Luciano Gomes Fietto.

In this work we have created an *in silico* system to address secretion of a desired protein among the genome data of *Kluyveromyces lactis*. The completed *K. lactis* genome sequencing has provided a tool to construct such a system. In order to explore a potential *K. lactis* extracellular secretome, four computational prediction algorithms have been applied: SignalP (presence or absence of an N-terminal signal peptide and clivage site), Phobius (transmembrane topology), big-PI Predictor (GPI modification site) and WolfPsort (subcellular addressing, including extracellular prediction). These algorithms have correctly predicted 95 yeast secreted proteins sought in public databases (NCBI, UNIProt and MIPS). They have also predicted as intracellular the same number (i.e. 95) of random sequences found in *K. lactis* database. The *K. lactis* database consists of 5327 sequences (http://cbi.labri.fr/Genolevures). When analyzed by SignalP 3.0, it has pointed out 698 putative proteins with N-terminal signal peptides. In this group, 260 were predicted by Phobius to have no transmembrane domains and 236 were found by the big-PI Predictor to have no GPI modifications site. Finally, the predicted *K. lactis* secretome was estimated to consist of up to 101 sequences by WolfPSORT which eliminates proteins with subcellular targeting. In order to validate theses analysis, both groups of predicted and annotated extracellular

proteins were compared by Hotelling's T^2 test. The analysis has shown no differences between the mean values of these two groups. The physiological significance of those potential extracellular proteins was similarly investigated by analyzing the relationship between the *S. cerevisiae* transcriptional regulators ortologues in *K. lactis* and the putative promoters (i.e. 1 KB upstream) of those extracellular proteins. It was applied the methodology proposed by "Yeastract" which search for elements such as binding sites that indicates associations between transcriptional factor and target genes. The physiological condition favoring protein expression in extracellular medium was obtained by searching Gene Ontology (http://www.geneontology.org). It has been shown that most of the transcriptional regulators of *K. lactis* extracellular proteins are related to stress response, especially presence of drugs into the medium. Also pH stress and limiting nitrogen can induce the extracellular proteins.

1. INTRODUÇÃO

Estima-se que até o final de 2008 mais de 500 genomas já tenham sido seqüenciados e que mais de 1100 estejam em andamento (http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome). Apesar dos genomas eucariotos ainda representaremrelativamente uma pequena porção desse total, cerca de 15%, cada genoma eucarioto gerade mais de 100 vezes do seu próprio volume dados em anotações e publicações. O genoma da levedura *S. cerevisiae,* que foi o primeiro genoma eucarioto seqüenciado e atualmente um dos mais bem estudados, é um genoma pequeno quando comparado aos dos eucariotos mais complexos econtribui com mais de 141 dados de processos biológicos, 180 banco de dados, 950 publicações, 2300 dados de microarray e milhões de interações entre proteínas e proteínas-pequenas moléculas(HUTTENHOWER et al., 2008).

O genoma da levedura *Kluyveromyces lactis*, que tem sido apontada como um dos modelos alternativos a *S. cerevisiae*, foi seqüenciado e anotado pelo consórcio *Genolevures* (http://cbi.labri.fr/Genolevures) em 2004. Análises subseqüentes mostraram relações filogenéticas próximas a *Saccharomyces cerevisiae* apesar de ter divergido anteriormente ao evento de duplicação do genoma (SHERMANet al., 2006). Embora apresentem uma alta relação na identidade das sequencias, diferenças fisiológicas marcantes têm sido mostradas para a *K. lactis*, tais como o metabolismo mais oxidativo, maiores rendimentos de biomassa, capacidade de assimilar diferentes fontes de carbono (por exemplo lactose), e pequena ou nenhuma repressão por glicose.

Atualmente, o grande desafio é entender como a diversidade fisiológica entre as duas leveduras é explicada apesar da similaridade genética. Estas diferenças residem em funções que determinam a comunicação da célula com o ambiente no qual ela se estabeleceu na natureza. Essas funções podem ser detectadas emproteínas presentes na via de secreção, na superfície celular, em vias de transdução de sinal ou no repertório de fatores transcricionais e seus respectivos sítio de ligação na região promotora.

A via de secreção em *K. lactis* desperta interesse pelo padrão de glicosilação das proteínas (LODI et al., 2005) o que, aliado ao metabolismo mais oxidativo e altos rendimentos de massa celular, favorece a produção de proteínas heterólogas. Proteínas extracelulares que passam pela via de secreção em levedura tem sinais em sua estrutura primária que determinam seu destino pós-tradução. Estes sinais consistem na seqüência de aminoácidos e na conformação do polipeptídeo que indicam se a proteína é substrato de uma enzima modificadora e/ou direcionam sua localização celular ou extracelular.

A pesquisa *in vivo* seria a melhor estratégia para entender as diversas funções contidas nas seqüências de nucleotídeos e aminoácidos depositadas nos bancos de dados, mas mesmo em organismos modelos como a *S. cerevisae* as diversas funções das proteínas ainda não são conhecidas nos diferentes níveis do metabolismo. O número cada vez maior de informações geradas pelos sequenciamentos dos genomas extrapola a capacidade das análises *in vitro* indicando a importância da análise preditiva ou simulações para avaliar as complexas relações biológicas. A bioinformática vem orientando as experimentações *in vivo* minimizando custos e otimizando o tempo, além de gerar hipóteses e propor novas questões estratégicas a serem testadas.

As técnicas moleculares requeridas em experimentos, tanto de ácidos nucléicos quanto de proteínas, dependem de informações consistentes e confiáveis (MEWES*et al*, 2000 e ALTSCHUL*et al,* 1997). Para isso são necessários grandes bancos de dados e algoritmos eficientes capazes de obter e armazenar as informações com o maior grau de precisão possível. A validação e a confiabilidade desses bancos de dados são muito importantes na mineração de resultados em diversos trabalhos experimentais (SEGRÈ, 2004).

Diante dessas considerações, o presente projeto propõe identificar *in silico* um conjunto de proteínas da levedura *Kluyveromyces lactis,* potencialmente secretadas para o meio extracelular e relacionar as seqüências de seus promotores putativos com os sítios de ligação dos fatores transcricionais documentados por BUSSEREAU *et al* (2006). Propõe-se ainda desenvolver um modelo booleano baseado na teoria dos grafos que possa mimetizar a rede regulatória celular que probabilisticamente estiver interagindo com as proteínas preditas como secretadas. Essa análise permitirá inferir através de operações padrão no modelo em grafos quais condições fisiológicas favorecem a máxima expressão e secreção de proteínas no meio extracelular.

CAPÍTULO 1

2. REVISÃO DE LITERATURA

Em meados da década de 40, o estudo dos genes envolvia o mapeamento individual e a análise de sua função e regulação. A partir do início da década de 70, surgiram novas técnicas genéticas de clonagem pela manipulação do DNA *in vitro*. Foi então possível estudar pequenas regiões de um genoma, inclusive regiões regulatórias e operons inteiros. Logo foi possível o seqüenciamento e a análise de genomas completos, o conjunto total de genes de um organismo. Em meados da década de 80, o termo genômica foi cunhado para descrever a ciência envolvida no mapeamento, seqüenciamento e análise de genomas (MADIGAN *et al,* 2004). O primeiro genoma celular seqüenciado foi o cromossomo de *Haemophilus influenzae*, contendo 1.830.137 pb (pares de base), sendo publicado em 1995 por Hamilton O. Smith, J. Craig Venter e colaboradores.O seqüenciamento e a análise da complexidade dos genomas não seriam possíveis se, paralelamente, não houvesse ocorrido melhorias nas tecnologias disponíveis. Por um lado ocorreu a automação do seqüenciamento. Por outro, o desenvolvimento e utilização de programas computacionais na análise, armazenamento e acesso das seqüências de DNA e de proteínas.

O primeiro genoma de eucarioto sequenciado foi o da levedura *Saccharomyces cerevisiae,* publicado em 1997, com aproximadamente 13.392

kb. Desde então foram publicados a seqüência do genoma da *Candida albicans* (TZUNG *et al*., 2001) e *Schizosaccharomyces pombe* (WOOD *et al,* 2001). Em 2004, o consórcio francês Genolevures publicou o seqüenciamento das leveduras *Candida glabata*, *Kluyveromyces lactis, Debaryomyces hansenii*e e *Yarrowia lipolytica* (SHERMAN *et al,* 2004).

Dentre estas últimas destaca-se a levedura *Kluyveromyces lactis*, que foi isolada de produtos lácteos e é reconhecida pela produção de β-galactosidase para hidrólise de lactose do leite e derivados e pela produção de biomassa e proteínas (Leonardo *et al,* 1991). *K. lactis* é uma levedura inócua e possui o *status* GRAS (*Generally Regarded as Safe*) reconhecido pelo Departamento de Agricultura dos Estados Unidos (*United States Department of Agriculture* – USDA). Recentemente, tem sido sugerida como hospedeira de genes heterólogos e, como modelo alternativo à *S. cerevisiae* em estudos sobre mecanismos moleculares ligados a doenças degenerativas associadas a mutações em genes mitocondriais (ROHOU et. al., 2001; RINALDI et. al., 2002). Além da assimilação da lactose, outra propriedade que distingue a *K.lactis* da *S. cerevisiae* é a ausência ou baixa repressão catabólicapor glicose. A *S. cerevisiae* é uma levedura tipicamente fermentativa e *Crabtree* positiva (GONZÁLEZ-SISO *et al*., 2000), ou seja, mesmo em altas concentrações de oxigênio dissolvido a fermentação predomina se a concentração de açúcar é alta. Em*K. lactis*, o metabolismo é predominantemente oxidativo e coexiste com o metabolismo oxidorredutivo ou fermentativo.A *K. lactis* também tem despertado grande interesse na área de secreção de proteínas heterólogas. Considerando que levedurasquando comparadas com os fungos filamentosos, não secretam naturalmente muitas proteínas, a etapa de purificação pode ser facilitadaem razão da ausência de contaminantes(BERG *et al*, 1990; BUCKHOLZ *et al*, 1991; OOYEN*et al*, 2006). Além disso, o padrão de glicosilação das proteínas secretadas por *K. lactis* é mais próximo a de eucariotos mais complexos (DOMINI*et al*, 2004).

2.1. Secreção de proteínas

A translocação de proteínas do citoplasma para a membrana e para o meio extracelular é vital para a maioria dos organismos, poisinfluencia a

resposta da célula a diversos sinais ambientais (EMANUELSSON, 2002). A via de secreção de proteínas é bem conservada em eucariotos ea informação necessária para oendereçamento aos diferentes locais celulares está contida na seqüência primária da proteína, tais como motivos ou seqüências de aminoácidos conservados.As proteínas que são destinadas àvia de exportação celular são sintetizadas em polirribossomos ligados às membranas do retículo endoplasmático rugoso (RE). As subunidades do ribossomo não especificam o polipeptídeo que está prestes a sintetizar, de modo que a iniciação e a elongação começam no citosol.

Proteínas secretadas têm um peptídeo sinal hidrofóbico próximo da extremidade aminoterminal. Não existe uma seqüência de peptídeo sinal específica, mas suas características incluem uma porção N-terminal com carga positiva, uma região central de 8-12 aminoácidos hidrofóbicos e um segmento C-terminal mais polar que, eventualmente, pode servir como ponto de clivagem para excisão do peptídeo sinal (BENDTSEN e*t al,* 2004).

O peptídeo sinal, com 15-30 resíduos de aminoácidos emerge do ribossomo no início da síntese do polipeptídeo. Ele liga-se a uma partícula de reconhecimento do sinal (SRP, *signal recognition particle*) na membrana do retículo endoplasmático. SRP é uma partícula alongada, constituída por seis proteínas diferentes e uma pequena molécula de RNA (7S), que serve de esqueleto. A ligação de SRP interrompe a síntese protéica e o complexo ribossomo e mRNA é direcionado ao retículo.

SRP reconhece e liga-se a um receptor de SRP ou proteína de aportagem (“docking”), localizada na superfície citosólica da membrana do RE, numa reação que requer hidrólise de GTP e, presumivelmente, envolve mudanças conformacionais no SRP e/ou no receptor. Este complexo (ribossomo + mRNA) é transferido a um “translocon”, um receptor na membrana que permite a passagem do polipeptideo que está sendo sintetizado (CHOU*et al*, 2004).Tanto SRP como a proteína de aportagem são liberados para dirigirem outros complexos ao RE, aliviando o bloqueio da tradução, causado por SRP. A seqüência sinal hidrofóbica, provavelmente complexada com um receptor protéico, é inserida na membrana, ancorando mais fortemente o ribossomo ao RE. A tradução e a extrusão na membrana, ou através dela, agora estão acopladas. As proteínas do translocon formam um poro ou um

canal através do qual o polipetídeo nascente é secretado; mesmo segmentos muitos hidrofóbicos ou iônicos são direcionados, através da membrana hidrofóbica, para dentro do lúmen do REe o enovelamento (*folding*) que gera as estruturas secundárias e terciárias se inicia (BRENDEN& TOOZE, 1999).

A proteína destinada à exportação completa fica, provavelmente, ancorada à membrana no lúmen do RE pelo peptídeo sinal. Um sitio de clivagem na proteína é hidrolisado pela peptidase sinal, uma proteína integral de membrana localizada na face luminal doRE. Então a proteína liberada começa o enovelamento, podendo formar pontes dissulfetos e em proteínas com múltiplas subunidades, estasiniciam a sua organização.

Existem ainda muitas outras variações que são encontradas nos peptídeos sinal. Existe uma classe particular que não é clivada que são as âncoras sinal, a qual fica presa na membrana durante a translocação (ALBERTS*et al,* 2002). O resultado é que toda proteína é ancorada na membrana. As âncoras diferenciam dos peptídeos sinal somente pela região do sítio de clivagem, em geral estes têm uma longa região *h*, tipicamente de mesmo comprimento da região transmembrana da α-hélice.Um outro tipo importante de associação com a membrana é mediada pela glicosilfosfatidilinositol (GPI). Uma proteína é primeiramente marcada para o RE pelo seu peptídeo sinal, onde a porção C-terminal é clivada, enquanto a âncora GPI é covalentemente ligada ao chamado sítio π (EMANUELSSEN, 2002).Outras etapas incluem o processamento proteolítico e a glicosilação, que ocorre no lúmen do RE e durante o trânsito da proteína pelo aparelho de Golgi e nas vesículas secretórias.

Proteínas em levedura são retidas no lúmen do RE em resposta a uma seqüência HDEL (His-Asp-Glu-Leu) C-terminal (HORTON *et al*, 2006). Quanto às proteínas que ficam retidas no Golgi, domínios transmembranas específicos e N-glicosilações mediados por glicosiltransferase tipo II (N-terminal citoplasmático e C-terminal lumenal) previnem o transporte de proteínas para o pós-golgi (YUAN*et al,* 2002). Nas proteínas endereçadas para o vacúolo o sinal de direcionamentoé caracterizado como um pró-peptideoque expõem o N-terminal após a clivagem da seqüência sinal. Contudo, para Nakai&Horton (2004), parece não haver motivos conservados, exceto por uma assinatura

(Thr/Ile/Lys) Leu Pro(Leu/Lys/Ile)N-terminal após o sitio de clivagem da peptidase sinal.

2.2. Fatores transcricionais

BUSSEREAU *et al*, 2006, utilizaram os motivos de ligação ao DNA:Dedos de Zinco Cys_2-His_2, Dedos de Zinco Cys_2-Cys_2, Dedos de Zinco Binucleares Cys_6, Zíperes de Leucina, bZIP (ziper de leucina de região básica) ebHLH (hélice-alça-hélice básico) para determinara presença dos reguladores transcricionais putativos de *K. lactis* por meio de comparações com os fatores da *S. cerevisiae*.Posteriormente,estes fatores foram agrupados pela presença destes motivos na sua estrutura. O dedo de zinco contém módulos similares de ~30 resíduos repetidos em *tandem*, cada um dos quais contendo dois resíduos de Cys invariáveis, dois resíduos de His invariáveis e vários resíduos hidrofóbicos conservados, sendo que cada uma dessas unidades liga-se a um íon de Zn^{2+}, que, conforme indicado por medições de absorção por raios X, fica tetraedricamente ligado aos resíduos invariáveis de Cys e His (KELLIS *et al*, 2003). As análises de seqüências revelaram que essas estruturas ocorrem de duas até mais de 60 vezes em diversos fatores de transcrição eucarióticos. Em alguns fatores, os dois resíduos de histidina que se ligam ao Zn^{2+} são substituídos por dois resíduos de Cys adicionais, enquanto outros possuem resíduos de Cys ligando dois íons Zn^{2+}. Em todos os casos, porém, os íons Zn^{2+} parecem unir domínios globulares relativamente pequenos, eliminando assim, a necessidade de núcleos hidrofóbicos muito maiores.

O zíper de leucina é um domínio estrutural que forma um dímero em espiral enrolada com repetições de Leu a cada sétima posição de um segmento de 28 resíduos do domínio. Repetições héptades similares ocorrem em algumas proteínas de ligação ao DNA de levedura tal como o Gcn4 (TODD& ANDRIANOPOULOS, 1997). Em extremidades N-terminal imediatamente adjacente ao zíper apresenta-se o chamado zíper de leucina de região básica (bZIP). Comparações de seqüências revelaram que a seqüência básica de 16 resíduos invariavelmente termina 7 resíduos antes do resíduo de Leu N-terminal do zíper de leucina. Além disso, toda essa região básica, bem

como o segmento de 6 resíduos que a une ao zíper de leucina, não possui os dois resíduos desestabilizadores de hélice mais fortes, Pro e Gly, o que sugere que cada polipeptídeo de bZIP é inteiramente α-helicoidal.Os motivos bHLH (*Basic Helix-loop-helix*) que ocorre em vários fatores de transcrição eucarióticos, contém uma região básica de ligação ao DNA conservada. Essa região é imediatamente seguida por duas hélices anfipáticas conectadas por uma alça que participada dimerização da proteína. O motivo de bHLH é, em muitas proteínas, seguido por um motivo de zíper de leucina conservado, que, presumidamente, aumenta a dimerização da proteína (Voet&Voet, 2006).

Utilizando essas propriedades moleculares dos reguladores transcricionais, HEINEMEYER *et al*, 1998, propuseram um sistema (http://www.cbrc.jp/research/db/TFSEARCH.html) que determina putativamente os motivos de ligação ao DNA nas seqüências de aminoácidos, e o correlaciona por meio dos elementos consenso dos sítios de ligação do DNA às possíveis seqüências "upstream" dos genes de interesse. TEIXEIRA*et al*, 2006,implementaram e otimizaram esse método para o caso específico da levedura *S. cerevisiae*, criando o sistema "Yeastract" (http://www.yeastract.com) que permite criar uma relação simples entre os reguladores preditos e documentados com os genes de interesse.

2.3. Algoritmos e Banco de Dados

Conforme proposto por Emanuelsson *et al*, 2007, os sistemas de predição levam em conta diversas características biológicas estruturais para acurar os algoritmos. A idéia é procurar essas regiões conservadas, tanto nas proteínas secretadas quanto nos fatores de transcrição, para conseguir determinar a probabilidade de certas sequências serem realmente codificadoras das proteínas de interesse *in vivo*. As abordagens mais usadas baseiam-se nos métodos de aprendizado de máquinas tais como as redes neurais e os Modelos Ocultos de Markov (HMM, *hidden Markov model*). Estes conseguem cobrir com o maior escopo possível os dados biológicos, devido ao seu forte caráter dinâmico, alto grau de complexidade e polimorfismos inerentes à variabilidade genética (BENDTSEN*et al*, 2004; MATSUDA*et al*, 2005).

Das redes neurais, a propriedade primordial é a habilidade de aprender a partir de seu ambiente e de melhorar o seu desempenho. Esta melhoria do desempenho ocorre com o tempo de acordo com alguma medida preestabelecida. Uma rede neural aprende acerca do seu ambiente por meio de um processo interativo de ajustes aplicados a seus pesos sinápticos e níveis de desvio. Idealmente, a rede se torna mais instruída sobre o seu ambiente após cada iteração do processo de aprendizagem (HAYYKIN, 2004). Vários algoritmos de predição de função e localização de proteínas usam esta técnica por exemplo o SignalP (BENDTSEN *et al*, 2004), PSORT que identifica peptideo sinal (NAKAI*et al*, 1999) e o wolfPSORT que determina motivos de endereçamento celular (HORTON*et al*, 2006).

Os Modelos Ocultos de Markov (Hidden Markov Models – HMM), são chamados de sistemas de aprendizagem não-supervisionados (sistemas auto-organizados) que se inspiram na mecânica estatística como fonte de idéias. Basicamente, representa o processo estocástico (putativo) gerado por uma cadeia de Markov subjacente, e um conjunto de distribuições de observações associadas com os seus estados ocultos (HAYYKIN, 2004). Esse modelo permite reconhecer padrões nas cadeias polipeptídicas. Os algoritmos aplicados a proteínas que usam esta técnica incluem o TMHMM que determina domínios transmembranas e âncora GPI (Krogh*et al,* 2001), TargetP que reconhece motivos e sinais de endereçamento de organelas (EMANUELSSON *et al*, 2000) e pTarget que compara as sequências de interesse com as armazenadas e as agrupam por funções celulares (GUGA, 2006).

Bussereau *et AL* (2006) identificaram os fatores de transcrição no genoma da *K. lactis*, alinhando as seqüências dos motivos de ligação ao DNA com aquelas conhecidas em *S. cerevisiae*. Os autores discutiram a relação entre os fatores identificados com a evolução e meio ambiente da*K. lactis*. Eles se restringiram à anotação,sem incluir outros componentes das vias de transdução de sinal ou proteínas de resposta. Ainda não se encontram publicados trabalhos que propõem uma relação de proteínas extracelulares com os fatores transcricionais de *K. lactis* que permitam sugerir uma condição de cultivo que resultasse numa máxima produção de proteínas extracelulares.

1.4. Teoria dos grafos e a rede regulatória

A teoria dos grafos é uma área sofisticada da matemática e ciência da computação que se desenvolveu nos últimos 200 anos, desde que foi primeiramente estudada. Muitos resultados são de interesse teórico, mas atualmente a sua aplicabilidade estende-se além da ciência da computação e é cada vez mais usado em sistemas naturais que exibem alta complexidade (BABU et al, 2004; HUTTENHOWER et al, 2008; CROMBACH et al, 2008).

Por definição, um grafo é uma estrutura de dados que é representada por uma coleção de vértices (ou nós) e as conexões entre eles (DROZDEK, 2002). Geralmente,nenhuma restrição é imposta ao número de vértices (V) ou ao número de conexões que um vértice pode ter com outros vértices. Portanto, os grafos são estruturas de dados versáteis que podem representar um grande número de situações e eventos diferentes, a partir de diversos domínios. Existem vários meios para se representar um grafo. Uma representação simples é dada por uma lista de adjacências, que especifica todos os vértices adjacentes a cada vértice do grafo. Outra representação é uma matriz que pode ser organizada em duas formas: uma matriz de adjacência e uma matriz de incidência (KLEIN et al, 2002). A matriz de adjacênciasdo grafo é uma matriz de dados binários Vx V tal que cada elemento dessa matriz

$$a_{ij} = \begin{cases} 1 \; se \; existe \; conexão \; (v_i v_j) \\ 0 \; caso \; contrário \end{cases}$$

A outra representação da matriz de um grafo é baseada na incidência de vértices e de arestas (E). Esta matriz de incidência do grafo é uma matriz VxE tal que

$$a_{ij} = \begin{cases} 1 \; se \; a \; aresta \; e_j \text{ é } incidente \; com \; o \; vértice \; v_i \\ 0 \; caso \; contrário \end{cases}$$

Podem ser utilizadas outras formas de representar os grafos além dos já citados,dependendo do problema em que esteja estudando.

Especificamente nas redes regulatórias todas essas formas podem ser utilizadas. Segundo BABU et al, 2004 a construção das interações regulatórias ligando fatores transcricionais e os genes alvo em um determinado organismo

pode ser visualizado por meio de um grafo direcional representado por uma matriz de adjacências. Os fatores transcricionais e os genes alvos são os vértices e as interações regulatórias as arestas. A figura 1 mostra como estas redes regulatórias são construídas pela perspectiva da teoria dos grafos. A rede resultante é um sistema complexo com múltiplos níveis. No modo mais básico a rede compreende num conjunto de fatores transcricionais, genes (ORFs) e sítios de ligação ao DNA (Figura 1a). No segundo nível, estas unidades básicas são organizadas em padrões recorrentes de interações chamados de motivos(Figura 1b). No terceiro nível os motivos agrupados em unidades transcricionais semi-independentes clamados de módulos (Figura 1c). E por fim, o nível mais alto que consiste na interação entre os módulos formando uma rede interconectada de alta complexidade (Figura 1d). As operações sobre esta rede requerem alto poder analítico de qualquer algoritmo usado, mas mesmo assim este modelo é altamente plástico e flexível para que possam ser criados modelos inferenciais relacionados a uma rede regulatória.

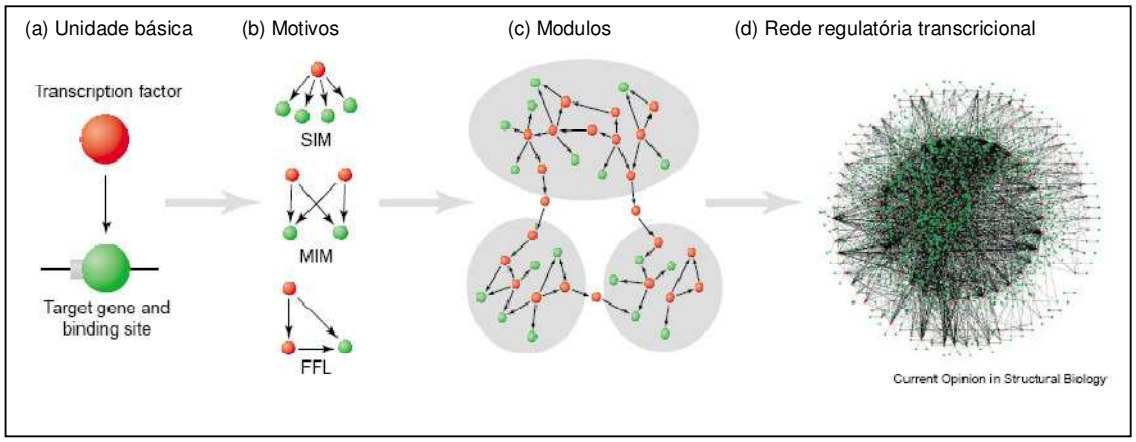


Figura 1 - Organização estrutural da rede regulatória transcricional. Desde a unidade básica (a) até o nível mais elevado de complexidade (d).

REFERÊNCIAS

ALBERTS, B.; JOHNSON, A. LEWIS, J.; RAFF, M.; ROBERTS, K.; WALTER, P. Molecular Biology of the Cell. Garland Science. New York, 2002.

ALTSCHUL, S.F.; GISH, W.; MILLER, W.; MYERS, E.W.; LIPMAN, DJ. Basic local alignment search tool. J MolBiol 215: 403-410, 1990.

ALTSCHUL, S.F.; MADDEN, T.L.; SCHÄFFER, A.A.; ZHANG, J.; ZHANG, Z.; MILLER, W.; LIPMAN, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research,Vol. 25, No. 17, 1997.

BABU, M.M.; LUSCOMBE,N.M.; ARAVIND,L.; GERSTEIN, M.; TEICHMANN, S.A. Structure and evolution of transcriptional regulatory networks.Current Opinion in Structural Biology, 14:283–291, 2004.

BENDTSEN, D.J.; NIELSEN, H.; HEIJNE, G.; BRUNAK, S. Improved prediction of signal peptides: SignalP 3.0.J. Mol. Biol., 340:783-795, 2004

BERG, J.A.; LAKEN, K.J.; OOYEN, A.J.J.; RENNIERS, T.C.H.M.; RIETVELD, K.; SCHAAP, A.; BRAKE, A.J.; BISHOP, R.J.; SCHULTZ, K.; MOYER, D.R.; SHUSTER, J.R.M. "*Kluyveromyces* as a Host for Heterologous Gene Expression: Expression and Secretion of Prochymosin". *Bio/Technology* 8, 135 - 139 (1990)

BOLOTIN-FUKUHARA, M.; TOFFANO-NIOCHE, C.; ARTIGUENAVE, F.; DUCHATEAU-NGUYEN, G.; LEMAIRE, M.; MARMEISSE, R.; MONTROCHER, R.; ROBERT, C.; TERMIER, M.; WINCKER, P.; WÉSOLOWSKI-LOUVEL, M. Genomic Exploration of the Hemiascomycetous Yeasts: 11. Kluyveromyces lactis. FEBS Letters 487 66-70, 2000.

BUCKHOLZ, R.G.; GLEESON, M.A.G. Yeast Systems for the Commercial production of Heterologous Proteins. *Bio/Technology* **9**, 1067 - 1072, 1991.

BUSSEREAU, F.; CASAREGOLA, S.; LAFAY, J-F.; BOLOTIN-FUKUHARA, M. The *Kluyveromyces lactis* repertoire of transcriptional regulators", FEMS Yeast Res 6,p.325–335, 2006.

CHEN, Y.; YU, P.; LUO, J.; JIANG, Y. Secreted protein prediction system

combining CJ-SPHMM, TMHMM, and PSORT, Mammalian Genome, Volume 14, Issue 12, p.859 – 865, 2003.

CHOU, K-C.; CAI, Y-D. Predicting protein localization in budding yeast. BIOINFORMATICS, Vol. 21, No.7, pages 944–950, 2005.

CROMBACH, A.; HOGEWEG, P. Evolution of Evolvability in Gene Regulatory Networks. PLoSComputBiol 4(7), 2008.

DONG, Q.; BALAKRISHNAN, R.; BINKLEY, G.; CHRISTIE, K.R.; COSTANZO, M.; DOLINSKI, K.; DWIGHT, S.S.; ENGEL, S.; FISK, D.G.; HIRSCHMAN, J.; HONG, E.L.; NASH, R.; ISSEL-TARVER, L.; SETHURAMAN, A.; THEESFELD, C.L.; WENG, S.; BOTSTEIN, D.; CHERRY, J.M. Gene Function, Metabolic Pathways and Comparative Genomics in Yeast. Proceedings of the 2003 IEEE Bioinformatics Conference (CSB'03).

DONNINI,C.; FARINA,F.; NEGLIA,B.; COMPAGNO,M.C.; UCCELLETTI,D.; GOFFRINI,P.; PALLESCHI, C. Improved production of heterologous proteins by a glucose repression-defective mutant of *Kluyveromyces lactis*. ApplEnviron Microbiol., 70(5):2632-8, 2003.

DROZDEK, A. Estrutura de dados e algoritmos em C++. Pioneira Thomson Learning, 2002.

EIDHAMMER, I.; JONASSEN, I.; TAYLOR, W.R. Protein Bioinformatics – An Algorithmic Approach to Sequence and Structure Analysis. John Wiley &Sons,Ltd. London, UK, 2004.

EMANUELSSON, O.; BRUNAK, S., HEIJNE, G.; NIELSEN, H. Locating proteins in the cell using TargetP, SignalP, and related tools. N*ature Protocols*2, 953-971, 2007.

EMANUELSSON, O.; NIELSEN, H.; BRUNAK, S.; HEIJNE, S. Predicting subcellular localization of proteins based on their N-terminal aminoacid sequence. J. Mol. Biol., 300: 1005-1016, 2000.

FISCHER, G.; ROCHA, E.P.C.; BRUNET, F.; VERGASSOLA, M.; DUJON, B. Highly Variable Rates of Genome Rearrangements between Hemiascomycetous Yeast Lineages. PLoS Genetics, Vol. 2, Issue 3, 2006.

GONZÁLEZ-SISO, M. I.; RAMIL, E.; CERDÁN, M. E.; FREIRE-PICOS, M. A. Respirofermentative metabolism in *Kluyveromyces lactis*: Ethanol production and the Crabtree effect. Enzyme and Microbial Technology 18:585-591, 1996.

GUDA, C. pTARGET: a web server for predicting protein subcellular localization. Nucleic Acids Research, Vol. 34, 2006.

HAIR, J.F., ANDERSON, R.E., TATHAM, R.L., BLACK, W.C. Análise Multivariada de Dados. Bookman, Porto Alegre, 2006.

HEINEMEYER, T.; WINGENDER, E.; REUTER, I.; HERMJAKOB, H.; KEL, A.; KEL, O.E.; IGNATIEVA, E.; ANANKO, O.; PODKOLODNAYA, F.; KOLPAKOV, N.; PODKOLODNY, N.; KOLCHANOV, J. DatabasesonTranscriptionalRegulation: TRANSFAC, TRRD, and COMPEL, NucleicAcids Res. vol.26, pp.364-370, 1998.

HIRSCHBERG, K.; MILLER, C.M.; ELLENBERG, J.; PRESLEY, J.F.; SIGGIA, E.D.; PHAIR, R.D.; LIPPINCOTT-SCHWARTZ, J. Kinetic Analysis of Secretory Protein Traffic and Characterization of Golgi to Plasma Membrane Transport Intermediates in Living Cells. The Journal of Cell Biology, Volume 143, Number 6, 1485-1503, 1998.

HORTON, P.; PARK, K-J.; OBAYASHI, T.; NAKAI, K. Protein Subcellular Localization Prediction with WoLF PSORT, Proceedings of *the 4th Annual Asia Pacific Bioinformatics Conference APBC06*, Taipei-Taiwan. p. 39-48, 2006.

HUTTENHOWER, C.; TROYANSKAYA, O.G. Assessing the functional structure of genomic data. Bioinformatics, vol. 24, pages i330–i338, 2008.

KELLIS, M.; PATTERSON, N.; ENDRIZZI, M.; BIRREN, B.; LANDER, E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements. NATURE, VOL. 423, 2003.

KLEE, E.W.; CARLSON, D. F.; FAHRENKRUG, S.C.; EKKER, S.C.; ELLIS, L.B.M. Identifying secretomes in people, pufferfish and pigs. Nucleic Acids Research, Vol. 32, No. 4, 2004.

KLEE, E.W.; ELLIS, L.B.M. “Evaluating eukaryotic secreted proteinprediction”**;** BMC Bioinformatics, **6**:256, 2005.

LEE, S. A.; WORMSLEY, S.; KAMOUN, S.; LEE, A.F.S.; JOINER, K.; WONG, B. An analysis of the *Candida albicans* genome database for soluble secreted proteins using computer-based prediction algorithms. Yeast; 20: 595–610, 2003.

LEONARDO, J.M.; BHAIRI, S.M.; DICKSON, R.C. Identification of upstream activator sequences that regulate induction of the beta-galactosidase gene in *Kluyveromyces lactis*. Mol Cell Biol.;7(12):4369-76, 1987.

LODI, T.; NEGLIA, B.; DONNINI, C. Secretion of Human Serum Albumin by Kluyveromyces lactis Overexpressing KlPDI1 and KlERO1. Applied and Environmental Microbiology, p. 4359–4363, 2005.

MADIGAN, T. M.; MARTINKO, J. M.; PARKER, J. Microbiologia de Brock. Pearson Prentice Hall, São Paulo/SP, 2004.

MATSUDA, S.; VERT, J-P.; SAIGO, H.; UEDA, N.; TOH, H., AKUTSU, T. A novel representation of protein sequences for prediction of subcellular location using support vector machines". Protein Sci. 14: 2804-2813, 2005.

MEWES , H.; FRISHMAN, W.D.; GÜLDENER, U.; MANNHAUPT, G.; MAYER, K.; MOKREJS, M.; MORGENSTERN, B.; MÜNSTERKÖTTER, M.; RUDD, S., WEIL, B. MIPS: a database for genomes and protein sequences**,** Nucleic Acids Res. 30: 31-34, 2002.

NAKAI, K.; HORTON, K. 'PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization', Trends in Biochemical Sciences, Volume 24, Issue 1, 1 January 1999, Pages 34-35.

NIELSEN, H.; KROGH, A.Prediction of signal peptides and signal anchors by a hidden Markov model. Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB 6), AAAI Press, Menlo Park, California, pp. 122-130, 1998.

NIKOLSKI, M.; SHERMAN, D.J. Family relationships: should consensus reign?—consensus clustering for protein families". BIOINFORMATICS, Vol. 23 ECCB, pages e71 – e76, 2006

OOYEN, A.J.J.; DEKKER, P.; HUANG, M.; OLSTHOORN, M.MA.; JACOBS, D.I.; COLUSSI, P.A.; TARON, C.H. 'Heterologous protein production in the yeast *Kluyveromyces lactis*', FEMS Yeast Research 6 (3), 381–392. (2006)

ROHOU, H.; FRANCISCI, S.; RINALDI, T.; FRONTALI, L.; BOLOTIN-FUKUHARA, M. Reintroduction of a characterized MittRNAglucine mutation into yeast mitochondria provides a new tool for the study of human neurodegenerative diseases.Yeast18(3):219-227, 2001.

SCHAFFRATH, R.; BREUNING, K.D. Genetics and molecular physiology of the yeast *Kluyveromyces lactis*. Fungal Genetics and Biology30, 173-190, 2000.

SCHAFFRATH,R.; MEINHARDT, F.; MEACOCK,P.A**.** Genetic manipulation of *Kluyveromyces lactis* linear DNA plasmids: gene targeting and plasmid shuffles. FEMS Microbiology LettersVol .178, 1999, Pages 201-210.

SEGRÈ, D. The regulatory software of cellular metabolism. TRENDS in Biotechnology Vol.22 No.6, 2004.

SHERMAN, D.; DURRENS, P.; BEYNE, E.; NIKOLSKI, M.; SOUCIET, J. Genolevures Consortium Génolevures: comparative genomics and molecular evolution of hemiascomycetous yeasts, Nucleic.Acids.Res. 32:D315-8, 2004.

SHERMAN, D.; DURRENS, P.; IRAGNE, F.; BEYNE, E.; NIKOLSKI, M.; SOUCIET, J. Génolevures complete genomes provide data and tools for

comparative genomics of hemiascomycetous yeasts. Nucleic Acids Research, Vol. 34, 2006.

STROUSTRUP, B. C++ Programming Language. AT&T Labs, Murray Hill, New Jersey. Addison-Wesley, 1997.

TEIXEIRA, M.C.; MONTEIRO, P.; JAIN, P.; TENREIRO, S.; FERNANDES, A.R.; MIRA, N.P.; ALENQUER, M.; FREITAS, A.T.; OLIVEIRA, A.L.; SÁ-CORREIA, I. The YEASTRACT database: a tool for theanalysisoftranscriptionregulatoryassociations in *Saccharomyces cerevisiae*. Nucleic Acids Research, Vol. 34, 2006.

TSAI, H-K.; CHOU, M-Y.; SHIH, C-H.; HUANG, T-W.; CHANG, T-H.; LI, W-H. MYBS: a comprehensive web server for mining transcription factor binding sites in yeast". Nucleic Acids Research, 1–4, 2007.

V. WOOD *et al.* The genome sequence of *Schizosaccharomycespombe*. *Nature***415**: 871-880, 2002.

VOET, D., VOET, J.G. Bioquímica; tradução Veiga, A.B.G. *et al*. 3.ed – Artmed, Porto Alegre, 2006.

WOLF, K., FWOLF, K. Nonconventional Yeasts in Biotechnology. 1.ed Springer, (2003)

YUAN, Z. TEASDALE, R.D. Prediction of Golgi Type II membrane proteins based on their transmembrane domains. Bioinformatics Vol.18, Num.8, Pag. 1109-1115, 2002.

CAPITULO 2

Computational analysis of the interaction between the transcriptional factors and the predicted secreted proteome of the yeast *Kluyveromyces lactis*

Brustolini, OJB[1]Fietto, LG[2], Cruz, CD[3], Passos, FML[1]

[1]Departamento de Microbiologia, Universidade Federal de Viçosa, Viçosa-MG, Brasil
[2]Departamento de Bioquímica, Universidade Federal de Viçosa, Viçosa-MG, Brasil
[3] Departamento de Biologia Geral, Universidade Federal de Viçosa, Viçosa-MG, Brasil

* ***Artigo submetido e aprovado no no periódico "BMC Bioinformatics"*** **e formatado segundo especificações desse periódico**.

ABSTRACT

In order to explore the potential *Kluyveromices lactis* extracellular secretome, four computational prediction algorithms have been applied to the 5327 *K. lactis* predicted proteins database. SignalP v3. has pointed out 698 proteins with N-terminal signal peptides. From those 260 were predicted to have no transmebrane domains by Phobius and 236 no GPI modification sites by big-PI Predictor. The predicted *K. lactis* secretome was estimated to consist of 101 putative proteins by WoLFf PSORT which excludes subcellular targeting. The transcription regulators of the putative extracellular proteins were investigated by searching the DNA binding sites on putative promoters. The conditions to favor expression were obtained by searching on Gene Ontology terms in database. Most of transcriptional regulators related to stress response such as presence of drugs, acid and heat, as well as low nitrogen concentration probably induce the maximum frequency of transcription of those potential proteins into the medium.

1. INTRODUCTION

The General Secretory Pathway (GSP) is a protein export process of major significance for both biological and technological reasons. Intercellular signaling and growth during development in multicellular organisms or cell communication depends on the secretion pathway. The export of a commercial protein into the extracellular medium by a recombinant cell would facilitate its downstream processing. The yeast *Kluyveromyces lactis* has been considered as promising host for heterologous protein production. Because yeasts naturally do not secrete as many proteins as filamentous fungi, it can support the production of secreted recombinant protein with few contaminants in the medium. An ideal system to address the secretion of a desired protein could be exploited among the native proteins. The completed *K. lactis* genome sequencing can provide the tools to construct such system. As the genomes of several hemiascomycetes yeasts are now sequenced (Tzung et al., 2001; Cliften et al., 2003; Kellis et al., 2003, 2004; Ramezani-Rad et al., 2003; Dietrich et al., 2004) and their gross comparison does not reveal great differences, the prospection for discovering a great potential secreted protein can be found applying the bioinformatics techniques.

In *K. lactis*, as in other eukaryotes, the secreted proteins are typically targeted by the presence of an N-terminal signal sequence to address them to GSP. Signal sequences usually have a well characterized structure composed by central hydrophobic core (h-region) and it consists an average of 6–15 amino acid (aa) residues which are flanked by hydrophilic N- and C-terminal regions. The h-region is important for correct targeting and membrane insertion of the peptide. The polar C-terminal region often contains helix breaking by the

occurrence of proline and glycine residues and small uncharged residues at the –3 and –1 positions which determine the signal peptide cleavage site. The polar N region is variable in length and frequently is positively charged (Bendtsen e*t al,* 2004). Although some proteins lacking N-terminal signal sequences reach the extracellular medium, the majority of soluble secreted proteins in *K. lactis* are likely to be transported via the GSP. A wide variety of computational methods have been assayed in order to predict the subcellular localization of proteins. The methods differ in terms of what input data they demand and what techniques are applied to make the decision (prediction) about location. Once the input data type are fixed, the methods for prediction-making are basically of two ways: the manually construction of the explicit rules for localization prediction from current knowledge of sorting signals, or applying data-driven machine learning techniques (eg Neural Networks or Hidden Markov Models, HMMs). The later automatically extracts decision rules from the sets of proteins with known location without making any prior detailed assumptions such as what features it is interesting to look for (Emanuelsson, 2002).

In addition to direct algorithms analysis to predict extracellular proteins, another approach can be made by relating extracellular secretome to its possible transcriptional factors. The transcriptional factors are one of the components of the transduction signal pathway which modulate the cell metabolism in response to environmental stimuli (Chekmenev *et al*, 2005). The transcriptional factors which contain DNA binding motifs are the closest component of the signaling pathway to DNA level. The combinatorial presence and absence of transcription factor binding sites (TFBSs) - to a large degree - is responsible for the gene regulation complexity in whole living organism (Wingender*et al*., 2000, 2001; Pickert*et al*., 1998; Kerl-Margoulis*et al*., 2003). The identification of TFBSs has been used to infer the regulatory networks for several yeasts (Bessereau*et al,* 2006).

By an algorithm approaching we proposed here to identify the extracellular protein candidates in the yeast *Kluyveromyces lactis* and its relation to TFBs. The analyses have consisted on evaluating the relationship between transcriptional regulators dataset published by Bussereau *et al,* 2006, and the putative promoter regions (1 KB upstream) of the genes coding the predicted extracellular proteins.

2. MATERIALS AND METHODS

2.1. Data Sets

The main data set analyzed in this work, was two files in FASTA format. Both files are composed by 5327 *K. lactis nucleotides and* amino acids sequences. These data are available in the *K. lactis* public second release from Génolevures consortium (http://cbi.labri.fr/Genolevures).

In order to test our extracellular proteins criteria, we assembled a validation set consisting of 95 non-redundant yeast extracellular proteins sequences (YEP) and 95 non-redundant *K. lactis* random sequences (KLRS). The YEP dataset was obtained by searching in the UniProt protein database (http://www.uniprot.org). The KLRS is assembled by using a random number generator and a sequence seeker algorithm.

Another validation dataset is manually extracted from the paper of Swain *et al*, 2008. It consists of 81 *K. lactis* extracellular proteins identified by Mass Spectrometry analysis (EPMS).

The *K. lactis* transcriptional factors (TFs) dataset used in this work was reported by Bussereau*et al,* 2006. The retrieved data are composed of 102 TFs identified as orthologues of *Saccharomyces cerevisiae* transactivators.

2.2. Algorithms and Strategy

We have applied the entire *K. lactis* predicted proteins dataset to SignalP v3.0 (http://www.cbs.dtu.dk/services/SignalP) to identify N-terminal signal peptides. With the intuit to define a positive SignalP hit, the following simultaneous criteria was pursued: (a) signal peptide predicted by SignalP Neural Network (NN) with the scores mean S and mean D; (b) signal peptide predicted by SignalP Hidden Markov Model (HMM) considering the value of probability and (c) signal peptide cleavage site located within 10-40 amino acid from the N-terminal.

The group of predicted ORFs which encode sequences with N-terminal signal peptides were then analyzed according to the presence of three additional characteristics: transmembrane domain, GPI modification site predicted by Phobius (http://phobius.sbc.su.se) and PI-predictor (http://mendel.imp.ac.at/gpi/gpi_server.html) respectively; the subcellular location was estimated by running the WoLF PSORT (http://wolfpsort.org) to identify signal addressing of subcellular location. The obtained dataset comprises all the sequences which deduced proteins are potentially acting in extracellular space. The outcome set was analyzed by PFam database (http://www.sanger.ac.uk/Software/Pfam) to update the '*Genolevures*' annotation.

In order to correlate the computational extracellular proteome and the transcriptional factors repertoire, we have created the first supporting algorithm based on C++ strings operations (Stroustrup, 2004) in *K. lactis* chromosomes dataset to retrieve 1 KB upstream sequence (putative promoter region) from each predicted extracellular ORF. The recovered dataset has been stored in FASTA file with the respective identification. The relationship between our computational extracellular proteome and transcriptional regulators repertoire was made according to Yeastract (Teixeira *et al,* 2006). The Yeastract web tools (http://www.yeastract.com) and database were used to find the associated transcriptional factors binding sites (TFBS) related to *S. cerevisiae*. We have

created the second supporting C++ algorithm to remove the *S. cerevisiae* transcriptional factors non-homologous to *K. lactis*.

2.3. Statistical Analysis

The multivariate analysis of variance was applied to verify the accuracy and determine the error rate of the computational secretome. The SignalP Neural Network scores (mean S and D) and SignalP Hidden Markov Model probability were used as values in statistical analysis which determine the matrices of variance-covariance of the predicted and validations sets. It was adopted the M-Box test for homogeneity of variance-covariance matrices between two groups, and Hotelling T^2 to calculate the probability of equality of the means vectors.

3. RESULTS

3.1. Prediction of *K. lactis* extracellular proteins

The flowchart of the algorithms used to seek the potential extracellular *K. lactis* proteins dataset is illustrated in Figure 1. By using standard criteria of the SignalP v3.0 we have predicted by NN and HMM scores from 5327 *K. lactis* ORFs, 698 ORFs containing consensus sequences for N-terminal signal peptides and signal peptidase cleavage site within 10-40 aa residues. Therefore, when the 698 deduced proteins harboring N-terminal signal peptides were submitted to Phobius algorithm only 438 were predicted to carry extra transmembrane domains, i.e. excluding the transmembrane domain of signal peptide. The following analyses were conducted with the remaining 260 ORFs. To identify GPI modification site the remaining ORFs were submitted to big-PI Predictor. The results indicated that 234 ORFs have contained the signal Peptide, no-transmembrane domain and no GPI modification site. Like some GSP proteins may be target to intracellular organelles rather than extracellular medium, we have used the algorithm WoLF PSORT to detect conserved addressing signatures to the organelles. The outcome has pointed out 162 ORFs predicted to extracellular addressing. Among them 101 ORFs have shown the highest *k*-NN score (~17.78 ± 5.68) and the others 61 ORFs have presented lower *k*-NN scores (~5.354 ± 3.989). The further analyses of the *K.*

lactis computational extracellular secretome were reasonable considered the 101 ORFs with the highest WoLFPsort *k*-NN scores.

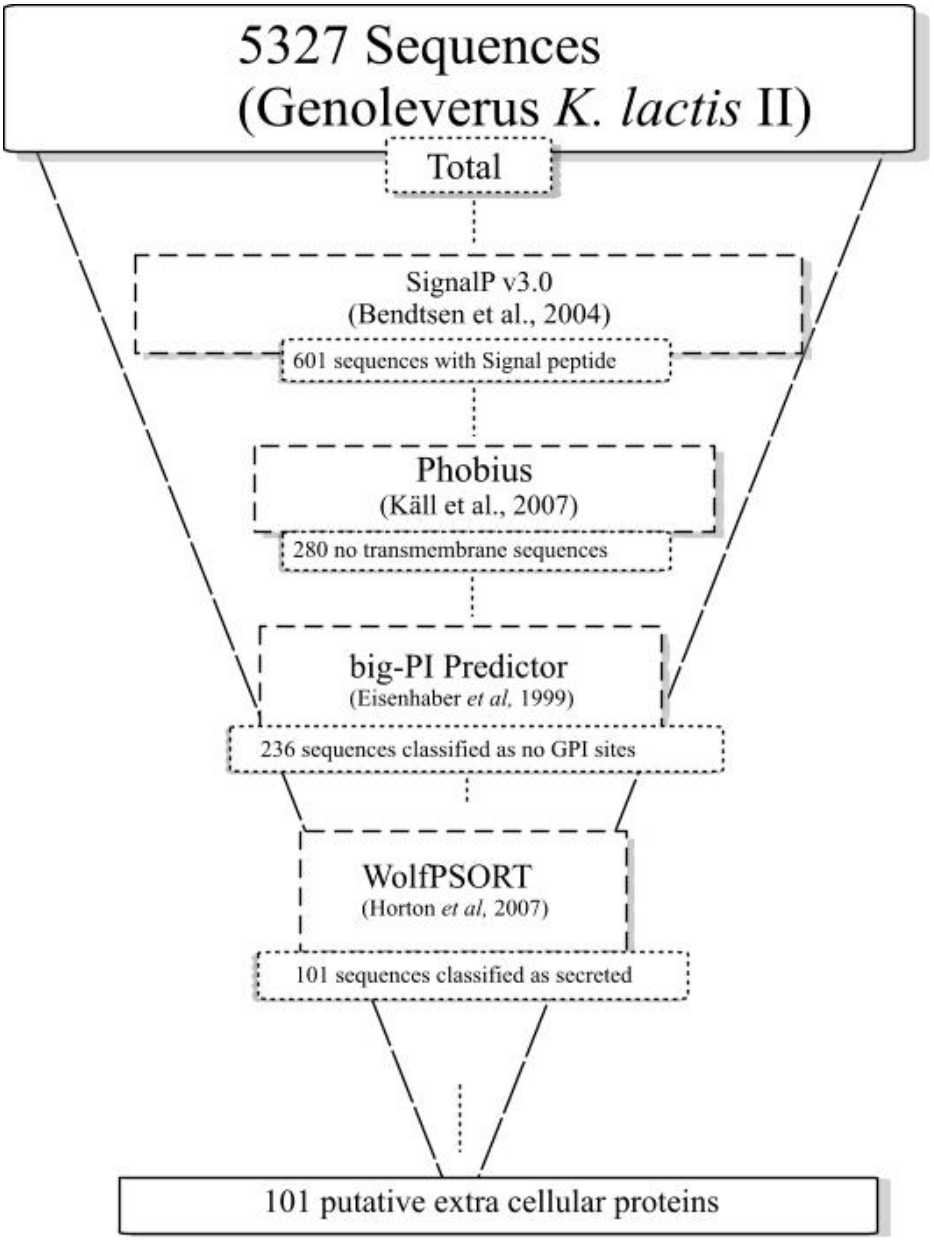
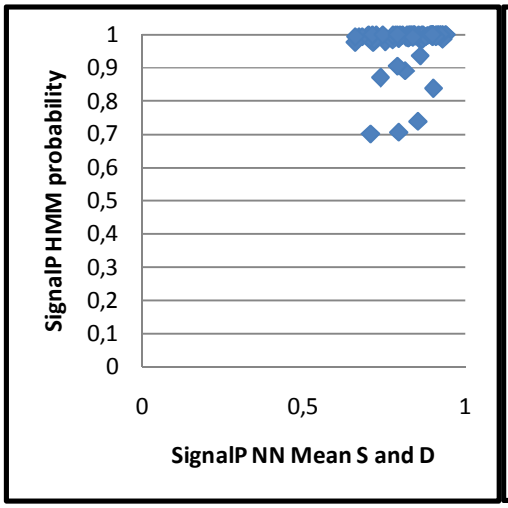


Figure 1 – Flowchart of strategy adopted to mining *K. lactis* gene sequences for extracellular proteins and respective outcoming results

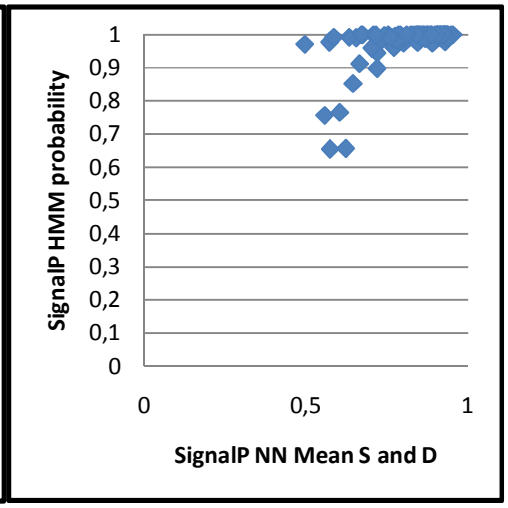
By a statistical approach we have tested the first GPS criteria, signal peptide, in the following datasets: yeasts extracellular proteins (YEP), *K. lactis* random protein sequences (KLRS) and the predicted extracellular proteins determined by WoLF PSORT (Figure 2). The YEP scores have shown the NN

S/D greater than 0.66 and the HMM around 0.8 while KLRS has presented these scores below 0.4 and 0.3 simultaneously (figure 2A and 2C). The comparison between the controls SignalP and the secreted ORFs scores have revealed that the scores of the 101 ORFs were very similar to the YEP, NN S/D 0.56 and HMM 0.78 (figure 2B). Therefore the standards criteria provided by SignalP were correctly encountered in all 95 sequences from the positive control. In order to evaluate the criteria for predicting the presence or absence of N-terminal signal peptides in *K. lactis* dataset, it was employed the Hotelling T-square multivariate test (Fig. 2D) based on NN Mean S/D and HMM score. The parameters vectors of each control set with the predicted set were compared and confirmed by T-square test. The estimated 101 ORFs are closer to the YEP dataset ($p=0.94$) than the KLRS ($p<0.001$).

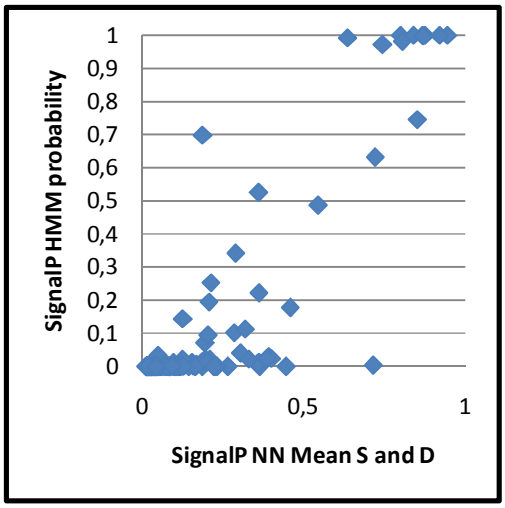
A



B



C



D

Hotelling T^2 Multivariate Test		
	T^2	P-value
Annotated vs Predicted group	0.1241	0.9401
Random vs Predicted group	927.243	<0.01

Figure 2 – Controls analysis by distribution of SignalP v3.0 scores (A) 95 yeast extracellular proteins (YEP) dataset; (B) 101 *K. lactis* predict extracellular proteins; (C) 95 *K. lactis* ramdom sequences (KLRS) from genome; (D)Multivariate tests using Hotelling T^2 to verify the

3.2. Analysis of annotations

The biological significance of the *K. lactis* predicted extracellular proteins by this work was based on the annotation available at 'Genolevures' web site. From the 101 predicted *K. lactis* extracellular proteins, 66 are annotated as similar to *S. cerevisiae*, 7to *Candida* genus and 4 with *K. lactis* documented proteins. The sequences were identified as complete or partial coding

sequences (CDS). More than 48% of known proteins are enzymes. A smaller group (4%) has predicted the pheromone or mating type function. Among the known sequences, 11% have been considered as intracellular proteins or wrong prediction (Fig. 3A). For those unknown potential *K. lactis* extracellular proteins (32%) we have applied the Protein Family Database (PFam) attempting to find any known conserved domains (Fig. 3B). The results have demonstrated that 9 singletons, among 21 singletons, harbor conserved domain with various PFam scores. The mating alpha factor precursor N-terminus (KLLA0A00154g, KLLA0F00220g), Kappa casein (KLLA0B05731g), NADH dehydrogenase subunit 2 C-terminus (KLLA0C10054g), Bacterial regulatory protein - Fis family (KLLA0D00660g), Thioredoxin (KLLA0E05544g), Mucin-like glycoprotein (KLLA0E10967g, KLLA0E19657g) and Collagen triple helix repeat (KLLA0F01595g) were the ones that presented a higher PFam scores. The analysis of the improbable secreted domains was shaped by alignment approach using BLAST tools (http://blast.ncbi.nlm.nih.gov/Blast.cgi). It was found that 6 from 9 sequences with non-secreted domains might have relation with extracellular proteins in other taxons (Data shown on our web site).

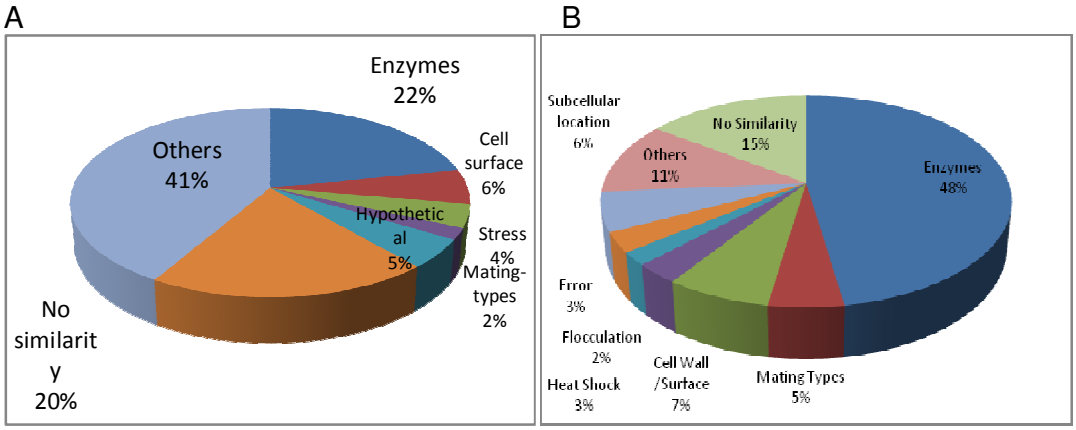


Figure 3 – Characterization of predicted proteins from (A) Génolevures II *K. lactis* annotation; (B) PFam for functional and conserved domains database (PFam).

3.3. Relationship between the predicted extracellular proteins and transcriptional factors repertoire

The 1KB upstream for each extracellular predicted ORFs, putative promoter region, was analyzed using Yeastract website tool to identify the TFBS related to *S. cerevisiae*. The results indicated the presence of 63 different transcriptional factors binding sites. In addition, in the *K. lactis* transcriptional factors dataset published by Bussereau *et al*, 2006, our supporting algorithm created to compared this dataset to *S. cerevisiae*, have found 27 TFs homologues in *K. lactis*(Table 1). We have established at least two TFBS in each promoter region in our analysis. In Yeastract database all the transcriptional factors have Gene Ontology terms (GO - http://www.geneontology.org), i.e. known detailed information about the cellular functional and addressing. These data have shown that all 101 sequences have the TFBS of Mot3p (involved in repression of a subset of hypoxic genes and repression of ergosterol biosynthetic genes), 96 of Fkh1p (a minor role in the expression of G2/M-specific transcription in mitotic cell cycle), 93 of Stb5p (Activator of multidrug resistance genes), 64 of Hac1p (regulates the unfolded protein response), 60 of Mcm1p (pheromone response), 57 of Gcn4p (activator of amino acid biosynthetic genes in response to amino acid starvation), 48 of Rgt1p (regulates expression of several glucose transporter (HXT) genes in response to glucose), 41 of Adr1p (peroxisomal protein genes, and of genes required for ethanol, glycerol, and fatty acid utilization), 41 of Nrg1p (mediates glucose repression and negatively regulates a variety of processes including filamentous growth and alkaline pH response), 41 of Pho4p (phosphorylation at multiple sites and by phosphate availability) and 37 of Yap1p (required for oxidative stress tolerance; mediates resistance to cadmium). These TFs dataset are the subgroup of our analysis that has been the probable major influence on the extracellular secretome.

Table 1. Summary of the computational analysis of the presence of TFBSs in putative promoter region of predicted extracellular ORFs and the related biological processes according to Yeastract /GO terms

Biological Process	*T.F.*	*ORFs*	*Description of Yeastract / GO terms*
Aerobic/ Anaerobic and Sterol metabolism	Mot3p	101	Repression of hypoxic genes, several DAN/TIR genes during aerobic growth, and ergosterol biosynthetic genes
	Ecm22p	2	Sterol regulatory element binding protein, regulates transcription of the sterol biosynthetic genes ERG2 and ERG3.
	Upc2p	1	Sterol regulatory element binding protein, induces transcription of sterol transport and biosynthetic genes; involved in the anaerobic induction of DAN/TIR mannoproteins and seripauperins
Cell Cycle	Fkh1p	96	The expression of G2/M phase genes; negatively regulates transcriptional elongation; positive role in chromatin silencing at HML and HMR.
	Cbf1p	29	Required for nucleosome positioning at this motif; targets Isw1p to DNA
	Ace2p	28	Activates expression of early G1-specific genes, localizes to daughter cell nuclei after cytokinesis and delays G1 progression in daughters.
	Rlm1p	7	Maintenance of cell integrity; phosphorylated and activated by the MAP-kinase
	Hcm1p	1	Drives S-phase specific expression of genes involved in chromosome segregation, spindle dynamics, and budding; telomere maintenance role
Drugs and metal resistance	Stb5p	93	Activator of multidrug resistance genes, forms a heterodimer with Pdr1p; interacts with a PDRE (pleotropic drug resistance element)
	Yap1p	37	Required for oxidative stress tolerance; activated by H2O2; mediates resistance to cadmium
	Yrr1p	18	Activates genes involved in multidrug resistance; paralog of Yrm1p, acting on an overlapping set of target genes
	Arr1p	5	Transcriptional activator required for transcription of genes involved in resistance to arsenic compounds
	Pdr1p	2	Master regulator involved in recruiting other zinc cluster proteins to pleiotropic drug response elements (PDREs) to fine tune the regulation of multidrug resistance genes
General stress response	Hac1p	64	Regulates the unfolded protein response, via UPRE binding, and membrane biogenesis; ER stress-induced splicing pathway utilizing Ire1p, Trl1p and Ada5p facilitates efficient Hac1p synthesis
	Gis1p	29	JmjC domain-containing histone demethylase; transcription factor involved in the expression of genes during nutrient limitation; also involved in the negative regulation of DPP1 and PHR1
	Sko1p	1	Forms a complex with Tup1p and Ssn6p to both activate and repress transcription; involved in osmotic and oxidative stress responses
	Msn2p	29	Transcriptional activator related to Msn4p; activated in stress conditions, which results in translocation from the cytoplasm to the nucleus; binds DNA at stress response elements of responsive genes, inducing gene expression
Pheromone response	Mcm1p	60	Involved in cell-type-specific transcription and pheromone response; plays a central role in the formation of both repressor and activator complexes.
Amino acid/ Nitrogen starvation response	Gcn4p	57	Amino acid biosynthetic genes in response to amino acid starvation; expression is tightly regulated at both the transcriptional and translational levels
	Met4p	20	Responsible for the regulation of the sulfur amino acid pathway, requires different combinations of the auxiliary factors Cbf1p, Met28p, Met31p and Met32p
	Lys14p	6	Involved in regulation of genes of the lysine biosynthesis pathway; requires 2-aminoadipate semialdehyde as co-inducer
Carbon source response	Rgt1p	48	Glucose-responsive transcription factor that regulates expression of several glucose transporter (HXT) genes in response to glucose; transcriptional activator and repressor
	Adr1p	41	Required for transcription of the glucose-repressed gene ADH2, of peroxisomal protein genes, and of genes required for ethanol, glycerol, and fatty acid utilization
	Azf1p	26	Involved in induction of CLN3 transcription in response to glucose; genetic and physical interactions indicate a possible role in mitochondrial transcription or genome maintenance
pH stress response	Nrg1p	41	Recruits the Cyc8p-Tup1p complex to promoters; mediates glucose repression and negatively regulates a variety of processes including filamentous growth and alkaline pH response
Phosphate response	Pho4p	41	Binds cooperatively with Pho2p to the PHO5 promoter; function is regulated by phosphorylation at multiple sites and by phosphate availability
Zinc (Zn) response	Zap1p	1	Zinc-regulated transcription factor, binds to zinc-responsive promoter elements to induce transcription of certain genes in the presence of zinc; regulates its own transcription; contains seven zinc-finger domains

The relationship between the transcriptional regulators and predicted extracellular proteome has shown a great complexity, then for attempting to create an *ab initio* model with biological significance, these relations have been shaped by graph theory approach. One of the graph representations is a square directed non-weight adjacency matrix. It has been created with 128 rows and columns. Among the 101 are our predicted proteins, 28 have their related transcriptional factors (TF). The graph was created by 128 nodes and 884 edges. As illustrate in Figure 5, the three sub-graphs have been extracted in order to illustrate the complexity of the regulatory network. We have used three well known extracellular proteins in *K. lactis,* α-factor mating pheromone (KLLA0E19173g), invertase (KLLA0A10417g)and acid phosphatase precursor (KLLA0A00176g). Our supporting material can be found in our WebSite: http://www.yeastmolphys.ufv.br/klactis.

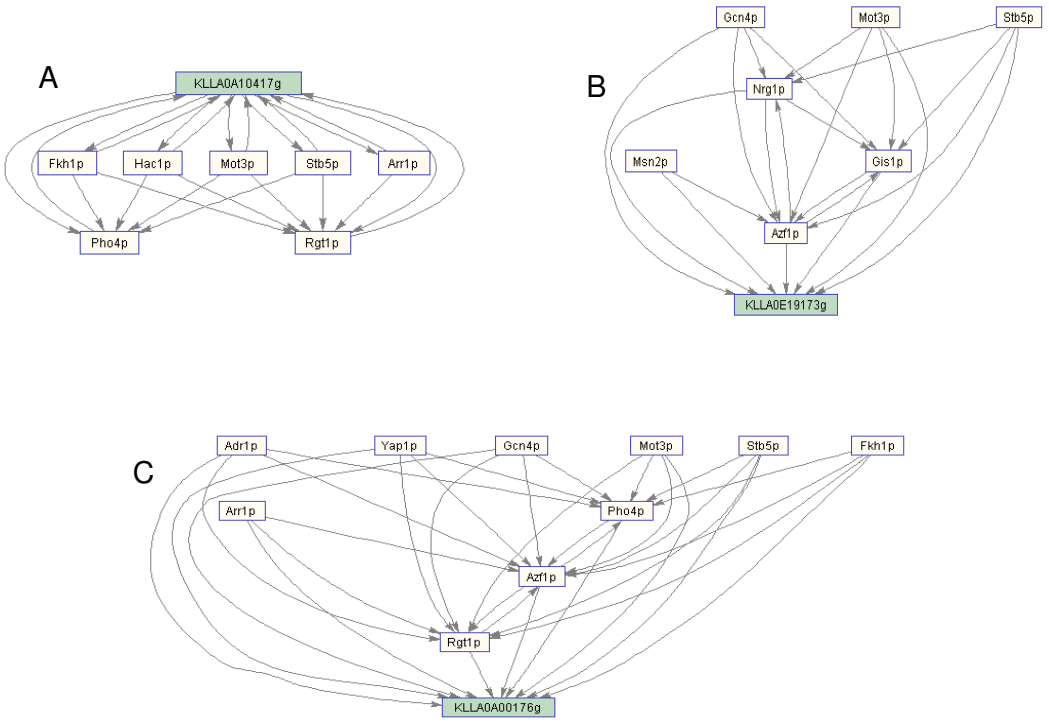


Figure 4 – Organization of the predicted transcriptional regulatory networks of (A) α-factor mating pheromone (KLLA0E19173g);(B) Invertase (KLLA0E19173); (C)acid phosphatase precursor (KLLA0A00176g).Transcriptional factors are represented by uncolored nodes and target ORFs by colored node. The edges are the presence of TFBSs in putative promoter region.

4. DISCUSSION

Considering its distinctive physiological properties, *Kluyveromyces lactis* has become an important model as non-*Saccharomyces* yeast. In addition, *K. lactis* has a great potential for biotechnological application including expression of heterologous proteins. These claims have motivated us to study the global extracellular proteome and correlate it to the transcriptional factors by using bioinformatics approach. Our final results have shown that a few numbers of proteins (101 proteins) are potentially secreted by *K. lactis*. In spite of using TMHMM and TargetP as used by Lee *et al,* 2003 and Swaim *et al,* 2008, we have applied Phobius and WoLF PSORT in the seeking step of transmembrane domains and subcellular addressing to detect targeted proteins to some organelles such as reticulum, golgi and proteasome. We have realized that the later algorithm seem to be more accurate, also when we compared the dataset of secreted proteins detected experimentally and published by Swain *et al,* 2008, the predicting methods of Lee *et al,* 2003 seem to detect a few more proteins (37) than the WoLF PSORT (33). However if we analyze the prediction error rate, the later method have shown about 69.3% and the other about 79.2%. As it is already known the occurrence of proteins in the medium can be changed according to different physiological conditions (Babu et *al*, 2004). Then, we have chosen the methods that can decrease the error rate in order to improve the obtaining a real extracellular protein in a given physiological condition. That error reduction lying in the incremented algorithmic Phobius (Käll*et al*, 2004) which may combine transmembrane topology and signal

peptide prediction and the new algorithm WoLF PSORT (Horton *et al,* 2007) which may predicts the subcellular localization site of proteins based on their amino acid sequences using *k*-NN (k-nearest neighbor) . As described by Swain *et al,* 2008, in the Signal Peptide detection step we have used the prediction algorithms SignalP v3.0 which gives two scores of Neural Network (NN) prediction: mean S and mean D and one score of Hidden Markov Model (HMM). These NN scores are used in the statistical analysis of the first necessary step of extracellular proteins from conserved secretory pathway: a signal peptide and signal peptidase cleavage site (Lee *et al*, 2003). When we analyzed extracellular proteins the accuracy might be decreased because the proteins that act in periplasmic space or cell wall pass through the general secretory pathway (GPS). It has not found yet motifs or conserved signal addressing for perisplasmic space or cell wall. Thus, the strategy adopted to classify our result was focused in annotation terms and in the PFam (database of collection of conserved domains and families). The Genolevures second release is the main public available annotation dataset for *K. lactis* sequences. Therefore, we were used the PFam database in addition to update the 'Genolevures' annotation. Both have shown that there are five *K. lactis* annotated secreted proteins: acid phosphatase, repressible acid phosphatase precursor, guanosine diphosphatase, exo-1,3-beta-glucanase and invertase. Although some of these proteins are not described to act in extracellular space, according to Domínguez *et al*, 1998, *S. cerevisiae* proteins are not found free in the extracellular medium but rather than they are retained in the periplasmic space or associated with the cell wall; on the other hand *K. lactis* does not seem to have the same characteristic, in fact it has been described to be able to secrete high molecular weight proteins (Becerra *et al*, 2001)**.** Thus, here we have considered proteins from periplasmic space or associated with cell wall as part of potential extracellular proteins dataset.

Like bioinformatics identifications are probabilistic in nature, their value lies in the low cost and high speed with which these identifications can be obtained; hence our analysis has continued exploit an *ab initio* model of physiological inference. We have created this model by using our computational extracellular proteome dataset, transcriptional regulators repertoire mining by Bussereau*et al,* 2006 and the Yeastract methodology created by Teixeiras *et al,*

2005 (http://www.yeastract.com). As gene expression programs depend on recognition of specific promoter sequences by transcriptional regulatory proteins (Lee *et al*, 2002), we decided to analyze the relationship between the Consensus sequences or DNA motifs binding sites with the transcriptional regulators. The content of many transcriptional regulators is one of the first changes occurring in cell after an environmental stimulus (Babu *et al*, 2004). When we have sought a set of *S. cerevisiae* transcriptional regulators ortologues and their related DNA motifs binding sites, it has been noticed that binding sites exhibit a high level of polymorphism (DNA binding factors capable of binding to both specific and nonspecific sequences). Due to the complex relation between transcriptional factors and predicted secretome we have shaped those data using Graph Theory (Patil and Nielsen, 2005). That empirical model may suggest many conditions which have not yet been thought by intuitive inference. The Gene Ontology terms described that each transcriptional factor dataset has shown the major of possible interactions related with stress and cellular cycle. Our results are in accordance with the literature description because expressions of extracellular proteins are increased in stress situation or in exponential phase where the cell needs proteins interacting in cell wall or in periplasmic space (Fermiñán and Domínguez, 2005). However for prospection of a good system to secretion it would be necessary to have a few different proteins in medium with high expression and secretion. With an *ab initio* model it is possible to seek the protein of interest and the possible environmental condition that might improve the expression and secretion. The model considers only the presence of transcription regulators binding sites in 1 KB upstream of predicted proteins. This is the first level among many more complexes. As any bioinformatics analysis, the graph theory modeling is dynamic and enhances the reliability. Our next goal will be test *in vivo* some models predictions we have made, as we have already found in the literature. This prospection will be useful to developing a specific system vector for heterologous proteins based on potential sequences of the predicted extracellular proteins.

5. REFERENCES

1. Becerra M, Prado SD, Siso MIG, &Cerdán ME (2001) New secretory strategies for Kluyveromyces lactis ß-galactosidase. Protein Engineering, Vol. 14, No. 5, 379-386.

2. Bendtsen DJ, Nielsen H, Heijne G &Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J MolBiol 340: 783-795.

3. Bolotin-Fukuhara M, Toffano-Nioche C, Artiguenave F, Duchateau-Nguyen G, Lemaire M, Marmeisse R, Montrocher R, Robert C, Termier M, Wincker P, Wésolowski-Louvel M (2000) Genomic Exploration of the Hemiascomycetous Yeasts: 11. Kluyveromyces lactis. FEBS Letters 487: 66-70.

4. Bussereau F, Casaregola S, Lafay JF&Bolotin-Fukuhara M (2006) The Kluyveromyces lactis repertoire of transcriptional regulators. FEMS Yeast Res. 6(3): 325-35.

5. Chen Y, Yu P, Luo J & Jiang Y (2003) Secreted protein prediction system combining CJ-SPHMM, TMHMM, and PSORT. Mammalian Genome 14(12): 859 – 865.

6. Chou K-C, Cai Y-D (2005) Predicting protein localization in budding yeast. BIOINFORMATICS 21 (7):944–950.

7. Eidhammer I, Jonassen I, Taylor WR (2004) Protein Bioinformatics – An Algorithmic Approach to Sequence and Structure Analysis. John Wiley &Sons,Ltd. London, UK.

8. Emanuelsson O, Brunak S, Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP, and related tools. N*ature Protocols*2: 953-971.

9. Emanuelsson O, Nielsen H, Brunak S &Heijne S (2000) Predicting subcellular localization of proteins based on their N-terminal aminoacid sequence. J MolBiol 300: 1005-1016.

10. Fischer G, Rocha EPC, Brunet F, Vergassola M, Dujon B. (2006) Highly Variable Rates of Genome Rearrangements between Hemiascomycetous Yeast Lineages. PLoS Genetics 2(3).

11. González-Siso MI, Ramil E, Cerdán ME, Freire-Picos MA (1996) Respirofermentative metabolism in *Kluyveromyces lactis*: Ethanol production and the Crabtree effect. Enzyme and Microbial Technology 18: 585-591.

12. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K (2007) WoLF PSORT: protein localization predictor. Nucleic Acids Res 35:W585-7.

13. Klee EW, Carlson DF, Fahrenkrug SC, Ekker SC, Ellis LBM (2004) Identifying secretomes in people, pufferfish and pigs. Nucleic Acids Research 32(4).

14. Lee SA, Wormsley S, Kamoun S, Lee AFS, Joiner K & Wong B (2003) An analysis of the Candida albicans genome database for soluble secreted proteins using computer-based prediction algorithms. Yeast 20: 595–610.

15. Matsuda S, Vert J-P, Saigo H, Ueda N, Toh H, Akutsu T (2005) A novel representation of protein sequences for prediction of subcellular location using support vector machines. Protein Sci. 14: 2804-2813.

16. Nielsen H, Krogh A (1998)Prediction of signal peptides and signal anchors by a hidden Markov model. Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB 6), AAAI Press, Menlo Park, California, pp. 122-130.

17. Nikolski M & Sherman DJ (2006) Family relationships: should consensus reign?—consensus clustering for protein families. BIOINFORMATICS 23: 71–76.

18. Schaffrath R, Breuning KD (2000) Genetics and molecular physiology of the yeast *Kluyveromyces lactis*. Fungal Genetics and Biology30: 173-190.

19. Segrè D (2004) The regulatory software of cellular metabolism. TRENDS in Biotechnology 22(6).

20. Sherman D, Durrens P, Beyne E, Nikolski M, Souciet J (2004) Genolevures Consortium Génolevures: comparative genomics and molecular evolution of hemiascomycetous yeasts. Nucleic Acids Res 32:315-8.

21. Sherman D, Durrens P, Iragne F, Beyne E, Nikolski M, Souciet J (2006) Génolevures complete genomes provide data and tools for comparative genomics of hemiascomycetous yeasts. Nucleic Acids Res 34.

22. Stroustrup B (1997) C++ Programming Language. AT&T Labs, Murray Hill, New Jersey. Addison-Wesley.

23. Swaim CL, Anton BP, Sharma SS, Taron CH, Benner JS (2008) Physical and computational analysis of the yeast Kluyveromyces lactis secreted proteome. Proteomics 8(13): 2714-2723.

24. Teixeira MC, Monteiro P, Jain P, Tenreiro S, Fernandes AR, Mira NP, Alenquer M, Freitas AT, Oliveira AL, Sá-Correia I (2006) The YEASTRACT database: a tool for theanalysisoftranscriptionregulatoryassociations in Saccharomyces cerevisiae. Nucleic Acids Res 34(Database issue):D446-51.

25. Wolf K, Fwolf K (2003) Nonconventional Yeasts in Biotechnology. 1.ed Springer.

26. Yuan Z, Teasdale RD (2002) Prediction of Golgi Type II membrane proteins based on their transmembrane domains. Bioinformatics 18(8): 1109-1115.

RESUMO E CONCLUSÕES

A secreção de proteínas é um processo de translocação celular de grande importância tanto por razões biológicas quanto tecnológicas. Por exemplo as vias de sinalização celular durante o desenvolvimento e o crescimento de organismos multicelulares, bem como a comunicação da célula com o seu meio, depende em grande parte da via de secreção. Do ponto de vista tecnológicoa secreção de uma proteína de interesse por uma célula recombinante e a sua posterior recuperação do meio de cultura desperta grande interesse comercial. A levedura *Kluyveromyces lactis* tem sido considerada como um modelo promissor para a produção de proteínas heterólogas, uma vez que leveduras naturalmente não secretam muitas proteínas como fungos filamentosos, oprocesso de recuperação da proteína é facilitadopela menor quantidade de proteínas contaminantes no meio. Um sistema ideal de secreção para uma proteína de interesse pode ser construído a partir dos elementos que influenciam o conjunto de proteínas nativas extracelulares.

Para identificar o secretoma extracelular putativo da *K. lactis*o uso dos algoritmos de predição SignalP v3 (peptideo sinal), Phobius (topologia transmembrana), big-PI predictor (ancora GPI) e WoLF PSORT (endereçamentos subcelulares) foram determinantes para agrupar as seqüências que possuíam as marcas para a via de secreção. Entretanto paraextrair dos dados uma significância biológica,o grupo de proteínas potencialmente secretadas foi submetida a uma atualização na anotação. Estes dados indicaram que a maioria das seqüências conhecidas (superior a 48%) pertenciam ao grupo das enzimas. Além da anotação foi verificado as interações das seqüências com a rede a regulatória transcricional por meio da presença de sítios de ligação ao DNA. Dados deste trabalho indicaram que os fatores transcricionais relacionados ao oxigênio, estresse e ciclo celular são os que provavelmente influenciam o maior número de proteínas secretadas. Como a regulatória transcricional exibe alto grau de complexidade foi utilizado o

modelo baseado na teoria dos grafos para armazenar e estudar estas possíveis relações. Estes permitem que possa ser determinado os módulos de regulação de cada proteína predita. Apesar do método ser bem estudado para as redes metabólicas, a sua relação com a regulação celular ainda está se iniciando. A perspectiva é que os dados obtidos do presente trabalho possam ser testados *in vivo* e o modelo inferencial sugerido possa ser validade ou melhorado.