UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Instituto de Biociências, Campus de Botucatu

ANALISANDO A DETERMINAÇÃO SEXUAL DE VERTEBRADOS COM BASE EM REDES DE INTERAÇÃO ENTRE PROTEÍNAS

GUILHERME TARGINO VALENTE

BOTUCATU - SP

2013

UNIVERSIDADE ESTADUAL PAULISTA

"Júlio de Mesquita Filho"

INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU

ANALISANDO A DETERMINAÇÃO SEXUAL DE VERTEBRADOS COM BASE EM REDES DE INTERAÇÃO ENTRE PROTEÍNAS

NOME DO CANDIDATO: GUILHERME TARGINO VALENTE

ORIENTADOR: CESAR MARTINS

CO-ORIENTADOR: NEY LEMKE

Tese apresentada ao Instituto de Biociências, Câmpus de Botucatu, UNESP, para obtenção do título de Doutor no Programa de Pós-Graduação em Ciências Biológicas (Genética).

BOTUCATU – SP

2013

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO DE AQUIS. E TRAT. DA INFORMAÇÃO
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CAMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ***ROSEMEIRE APARECIDA VICENTE***

Valente, Guilherme Targino.
Analisando a determinação sexual de vertebrados com base em redes de interação entre proteínas / Guilherme Targino Valente. – Botucatu : [s.n.], 2013

Tese (doutorado) - Universidade Estadual Paulista, Instituto de Biociências de Botucatu
Orientador: Cesar Martins
Co-orientador: Ney Lemke
Capes: 20204000

1. Genética. 2. Bioinformática. 3. Interação proteína - proteína. 4. Sistemas biológicos. 5. Evolução (Biologia). 6. Vertebrado – Evolução. 7. Sexo – Causa e determinação.

Palavras-chave: Genética; Bioinformática; Aprendizado de máquina; Redes complexas; Proteínas; Evolução; Determinação sexual.

Agradecimentos

A minha esposa Tatiana de Campos Melo, que, a partir do momento que entrou em minha vida, tem conduzido modificações incríveis na minha forma de agir, pensar e sentir. Apesar de sempre ter sido extrovertido e de tentar esconder como eu me sentia de verdade, antes de você eu acreditava que as coisas boas eram para todos, menos para mim; acreditava que os momentos bons eram passageiros e que nunca conseguiria ser ou a ter o que queria; admirava/acreditava nos outros mas nunca em mim; era mais ingênuo e "romântico" com a vida e pessoas; tinha poucos e pequenos sonhos, prazeres e planos; acreditava que jamais teria a oportunidade de te conhecer e muito menos de fazer alguém como você me amar. Hoje, no entanto, tenho tudo o que sempre sonhei e as coisas que não conquistei ainda não o fiz meramente por questão de tempo; batalho muito mais, sonho muito mais alto, acredito muito mais em mim e sei que vou conseguir; acredito que posso abraçar tudo, fazer tudo e ser tudo o que quero! Devo tudo isso a você e tenho muito orgulho do seu caráter, profissionalismo e de ter segurado a "barra" que foi esse péssimo início de 2013 (isso nos fortaleceu muito mais!). Te amo!

Aos meus pais, que me amam incondicionalmente e que apesar de todas as dificuldades e muitas vezes infelicidades, tentaram propiciar o máximo de carinho, atenção e incentivo para que nos desenvolvêssemos como cidadãos conscientes de nossos direitos e deveres e pessoas com sonhos e planos para conquistá-los (o pai sempre dizia: - Plano, programa, projeto!). Apesar da relativa "infeliz" adolescência e início de juventude que tive, conseguimos chegar até aqui. Espero que tenham orgulho do que tenho conquistado e que possamos aproveitar ao máximo tudo que o futuro nos guarda. Sintam-se como atores muito importante para a concretização desse trabalho. Sinto orgulho de vocês e também os amo!

Ao meu querido irmão, Gustavo Targino Valente, pela sua amizade incondicional, por ser meu melhor amigo e por ser esse "cara" espirituoso, divertido e com o poder de deixar todos ao seu redor confortáveis e amados. Você, apesar de ser mais novo, é para mim um grande exemplo de simplicidade (sem ser ingênuo), humildade, confiança, paz, profissionalismo e divertimento. Desejo que tenha uma vida excelente e que possamos estar sempre juntos! Te amo e tenho orgulho de você!

Ao meu orientador Cesar Martins, que ao longo desses 6 anos de convivência propiciou modificações incríveis na minha vida. Sair de Rio Brilhante (MS) e passar a viver em Botucatu (SP) foi a atitude mais enriquecedora da minha vida. Propiciou que eu pudesse explorar novos horizontes, construir planos, ser alguém admirado e propiciou um futuro que eu jamais imaginei. Possibilitou: aprender a fazer ciência; aprender genética (quando cheguei nem sabia que PCR existia); conhecer um congresso grande; o primeiro contato com uma Universidade de verdade; a primeira visita ao mar; a primeira viajem aérea; aprender inglês; a primeira cerveja importada; o primeiro restaurante elegante; a primeira viagem internacional; a primeira publicação; incontáveis viagens pelo país (AM, PA, PE, MT, MG, SP, RJ, PR e SC), coletas inesquecíveis, ... Além disso, muito de você carrego comigo: lições de convivência profissional (ensinando-me os "bastidores" da ciência, um privilégio de poucos pós-graduandos), profissionalismo, pró-atividade, eficiência ... Você foi fundamental para o desenvolvimento desse trabalho e da minha personalidade profissional.

Ao meu co-orientador Ney Lemke, figura extrovertida e divertida, que muito me ensinou ao longo desse doutorado. Seu papel foi fundamental para o meu desenvolvimento científico e conclusão dessa tese. Aprendi muito com você, aliás, muito mais do que imaginei.

Aos meus amigos Diogo Cavalcante Cabral-de-Mello e Rafal Takahiro Nakajima, por terem sido colegas de laboratório que se tornaram verdadeiros amigos. Sempre nos relacionamos de forma produtiva e divertida. Apesar de sermos muito diferentes, "no final das contas", nos parecemos em muitos aspectos. Aprendi muito com vocês e espero sempre ter o prazer de sermos amigos.

As minhas cunhadas Talita e Joseane, por trazerem felicidade as pessoas que mais amo.

Ao meu amigo Manolo Sanchez, o "índio" mais *playboy* do mundo! Figura esperta e companheira, que sempre me apoiou e mostrou-se interessado pelas coisas da minha vida. Apesar de não ter sido "padrinho" do meu casório por questões logísticas, sinta-se como um deles. Nossos pais eram amigos, somos amigos e espero que nossos filhos também sejam amigos. Abraços!

Aos meus tios Edmir e Margarete, que no início na minha vida Botucatuense me auxiliaram muito.

A professora Margarete, por ter sido uma excelente professora de inglês e por compreender que *homework* não é para mim! Sempre se divertia com a minha frase: - Nos países de língua inglesa, os pais dão *homework* para as crianças aprenderem inglês? Quero aprender como se estivesse lá!

Aos meus colegas de laboratório (incluindo ex-membros) e de outras instituições, Irani, Andréia, Juju Mazzucks, Jacú, Diogo, Sarah, Érica, Taka, Bruno, Édão, Marcão, Spike, Diego, Adauto, Marcio, Carlos Schneider, Claudia Gross, Leandra Terencio e Bianca, pelas conversas, discussões científicas, cervejas, baladas (*no passado tá amor!*), churrascos, caronas, brigas, por citarem meus artigos e por produzirmos ciência. Em especial ao Marcão, Taka, Marcio e Érica, por terem contribuído com algumas análises desse trabalho. Aprendi com cada um de vocês.

Aos amigos de república, Kékú, Xurupito, Chapeleta, Jacú, Caruncho, Roseana (forrózero), Parrudo, Édão e Juan (não vai a utilizar? Passa para mi?), pelas farras, diversões, companheirismo, caronas e as vezes até desacordos. Sinto saudades!

A Adriane Pinto Wasko, por me apresentar o programa "Experimentando Genética", o qual participei com muita vontade e carregarei comigo durante a minha vida profissional.

Aos colegas da University of Maryland, Thomas, Matt, Karen e Clair, por terem propiciado um excelente aprendizado científico, em especial o professor Thomas, que me deu a oportunidade de conhecer o *Broad Institute* e participar de uma conferência com vários pesquisadores importantes na área de evolução de ciclídeos.

Aos colegas da University of Connecticut, Rachel e Andrew. Pessoas brilhantes e simpáticas, que possibilitaram um ótimo aprendizado profissional.

Aos amigos Sérgio e Márcia, pessoas que muito me ajudaram e me trouxeram conforto quando precisei.

A dona Terezinha, Iolanda e Vanda, por deixarem o departamento sempre "limpinho" e confortável para que todos possam desenvolver os trabalhos com muito mais vontade e facilidade.

Aos amigos de república, Javy, Carlos, Liza, Donald e Shany, por possibilitarem que eu me "sentisse em casa" durante a minha estadia nos USA; e por me ajudarem com o inglês.

Ao técnico Zé Eduardo, por ter sido uma pessoa que muito me encorajou a estar nessa vida Botucatuense.

A Luciana, secretária do departamento de Morfologia, que, com muita maestria, “bota lenha nessa fogueira” e faz esse departamento “andar nos trilhos”.

Aos funcionários da pós-graduação, que sempre são hábeis e prestativos.

A pós-graduação em genética, por possibilitar condições muito boas para o desenvolvimento dos pós-graduandos.

Ao departamento de Morfologia, ao instituto de Biociência e a UNESP de Botucatu, pela infraestrutura fornecida para o desenvolvimento desse trabalho.

Aos professores membros da banca de qualificação e de defesa, por terem dispendido um precioso tempo contribuindo para essa tese. Assim, considerem-se agentes ativos na minha formação profissional.

A FAPESP e CNPq pelo auxílio financeiro, imprescindível para o desenvolvimento desse trabalho.

Ganhar? Não sei. Desistir? Nem tão fácil

Muito obrigado!

Índice

1. Resumo

Atualmente os sistemas biológicos vem sendo abordados por diversas áreas de pesquisas, dentre elas a biologia de sistemas. Essa área tem centrado esforços para descrever e compreender as relações entre os componentes bióticos e abióticos desses sistemas, utilizando para isso a teoria de grafos. Uma forma de estudar esses sistemas, é avaliar as interações entre proteínas, que são as interações biomoleculares mais abundantes nas células. Nesse contexto, a presente tese focou no desenvolvimento de um algoritmo computacional capaz de predizer interações entre proteínas de qualquer espécie ou conjunto protéico. Para isso, foram utilizadas técnicas de aprendizado de máquina (sub-área da inteligência artificial) para a construção e aplicação desse preditor, o qual mostrou-se eficaz em predizer interações entre proteínas de mais de 80 espécies diferentes, incluindo interações entre proteínas de parasita e hospedeiro. Esse novo preditor foi aplicado no proteoma de zebrafish (*Danio rerio*) e humanos (*Homo sapiens*), gerando assim redes de interações proteicas para ambas as espécies. As interações obtidas foram avaliadas em um contexto geral e o foco das análises foi direcionado para a sub-rede relacionada com as vias de determinação e diferenciação sexual em vertebrados. Os resultados demonstraram uma relativa baixa conservação desses grafos ao longo da evolução dos vertebrados, tanto do ponto de vista global quanto da sub-rede relacionada com a determinação e diferenciação sexual em vertebrados. Além disso, foi possível observar ao menos um *hub* conservado entre as duas sub-redes, o qual representa um novo alvo a ser avaliado por pesquisas experimentais. Contudo, os dados demonstram que o preditor gerado possui um grande potencial para diversas áreas de estudos e é bastante útil para predição de interações em larga-escala. Além disso, os aspectos evolutivos aqui estudados abrem novas perspectives acerca da evolução dos sistemas biológicos e das vias de determinação e diferenciação sexual em vertebrados.

2. Abstract

Nowadays the biological systems have been analyzed under several research foci, including the system biology. This area focus to describe and understand the relationship between biotic and abiotic factors using the graph theory. The protein interactions are target interactions to the system biology because they are the most abundant biomolecular interactions within a cell. Thus, this thesis reported the development of a computational algorithm to predict protein-protein interactions for all species or protein sets. The machine learning technique (sub-area of artificial intelligence) were used to develop and apply this method, giving effective results to predict protein-protein interaction for more than 80 different species, including parasite-host associations. This new predictor was applied to the proteome set of zebrafish (*Danio rerio*) and humans (*Homo sapiens*), generating the protein-protein interactions for both species. Evolutionary aspects of the protein interactions were studied in a broad context and the focus was directed to the sub-network involved in the vertebrate sex determination and differentiation. The results reported a low conservation of those graphs across the evolution in a general view or for the sub-network related to the vertebrate sex determination and differentiation. Moreover, it was reported at least one conserved hub between both sub-networks, to be further evaluated by experimental procedures. Anyway, the data showed that the predictor here reported may be very useful for several research areas and it is desirable for large-scale prediction procedures. Furthermore, the evolutionary aspects discussed in this thesis open new perspectives concerning system biology and evolutionary pathways of vertebrate sex determination and differentiation.

3. Introdução

A presente tese é focada no estudo biológico da determinação e diferenciação sexual de vertebrados, utilizando espécies de vertebrados como modelos experimentais. Esse tema foi abordado utilizando análises de redes complexas, mais precisamente no estudo de redes de interação entre proteínas. Nesse contexto, a introdução dessa tese está dividida em dois tópicos principais: 1 – o uso de redes complexas aplicados a estudos biológicos, visto que o foco do trabalho trata de um sistema complexo; 2 - o sistema de determinação e diferenciação sexual em vertebrados, aspecto biológico sob investigação no presente trabalho.

3.1 Redes biológicas

Ao longo dos anos os sistemas biológicos vem sendo analisados de forma fragmentada e reducionista e, apesar dessas análises serem de extrema importância, Junker (2008) descreveu que: "o ambiente é uma combinação de sistemas complexos interligados em vários níveis... ". Dessa forma, a necessidade e interesse de se estudar os sistemas vivos de forma holística vem aumentando cada vez mais. Essas abordagens tem incluído em sua plenitude estudos multi-escala (de um ponto de vista molecular a global), os quais buscam integrar todos os fatores bióticos e abióticos a fim de se compreender a dinâmica e a complexidade desses sistemas em um nível mais elevado.

Os sistemas biológicos possuem características como: 1 – desenvolver estratégias dependentes do sistema que os circunda; 2 – as entidades são heterogêneas e "lutam" contra o ambiente para manterem-se fora de um equilíbrio; 3 – interagem com o ambiente circundante modificando-o de acordo com as estratégias aplicadas para uma manutenção fora-do-equilíbrio. Além disso, aprendem com as próprias experiências recebendo um *feedback* do ambiente; 4 – a complexidade das entidades é relacionada a um grande número de variáveis bióticas e abióticas. Conceitualmente, qualquer sistema complexo é constituído por um grande número de entidades (partículas ativas ou agentes) que podem interagir mutuamente e com o ambiente exterior (Bellomo e Carbonaro, 2011). Dessa forma, todas as características de um sistema biológico são responsáveis ou produtos de uma complexa interação entre entidades de um sistema complexo. Por fim, desde uma partícula viral a um bioma inteiro, incluindo seus aspectos evolutivos e suas relações com os componentes bióticos a abióticos, podem ser então enquadradas dentro do conceito de sistemas complexos.

Teorias matemáticas vem sendo aplicada para estudos de sistemas biológicos em uma grande variedade de questões, tais como para a compreensão dos mecanismos moleculares e celulares, produção de fenótipos, função de um sistema biológico, aspectos comportamentais, populacionais, ecológicos e evolutivos (alguns desses fatores estão citados em Bellomo e Carbonaro, 2011). Não obstante, uma subárea da matemática conhecida como teoria de grafos (tendo a palavra "rede" como um termo informal) vem sendo aplicada nas teorias de sistemas biológicos, em estudos de redes bioquímicas, regulação transcricional, redes metabólicas, de transdução de sinais e de interação entre proteínas, com o intuito de responder questões de biologia funcional e estrutural, evolutivas e clínicas (Pavlopoulos et al. 2011).

A utilização de teoria de grafos para estudar aspectos biológicos é enquadrada dentro do que se conhece como biologia de sistemas. O principal objetivo dessa área é elucidar, modelar e predizer todos os comportamentos das interações de um sistema biológico. Como já reportado, é de extrema importância a incorporação de uma visão holística sobre o sistema biológico, e nesse sentido, a biologia de sistemas vem cada vez mais satisfazendo essas necessidades, tanto em nível micro como macroscópico. Nessas redes, todos os elementos bióticos e/ou abióticos são representados por nós ou vértices (inglês *vertice*, *nodes* ou *points*) e a relação entre eles por linhas, que são também conhecidas como arestas (inglês *edges*, *arcs* ou *lines*) (Junker, 2008) (Figura 1).

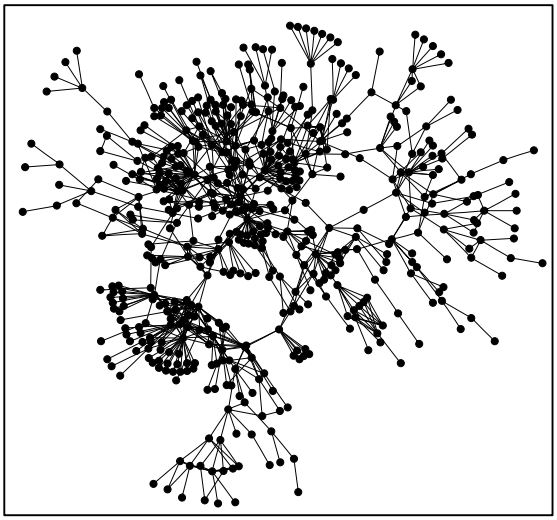


Figura 1 – Exemplo de grafo. Nesse exemplo, os nós e as linhas (arestas) podem ser encarados como proteínas e suas relações, respectivamente (Junker, 2008).

As redes de regulação reportam o controle da expressão gênica. A expressão final de um gene pode ser modulada por diversas variáveis tais como, fatores de transcrição, modificações pós-traducionais ou por associação dessas proteínas com outras biomoléculas. Nesses modelos, a rede é demonstrada por um grafo direcionado (Junker, 2008; Pavlopoulos et al., 2011) tal como V2→V3, significando que a molécula "V2" interage com "V3" e não o oposto. Em termos práticos, V2→V3 simbolizariam a interação unidirecional entre um fator de transcrição e o promotor de um gene, respectivamente (Figura 2).

As redes de transdução de sinais são representadas por grafos direcionados multi-arestas que representam as interações uni ou bi-direcionais entre proteínas, outras moléculas orgânicas ou inorgânicas. Em suma, essas redes reportam como o sinal extracelular pode ser transmitido para o ambiente intra-orgânico ou como os sinais são conduzidos dentro de uma célula ou organismo (Junker, 2008; Pavlopoulos et al. 2011) (Figura 2).

As interações entre proteínas são geralmente reportadas como grafos não-direcionados uma vez que não importa a direção que a interação ocorre (Figura 2) (Junker, 2008). O presente trabalho é baseado nesse tipo de rede, sendo os capítulos 1 e 2 dedicados a esse tipo de interação.

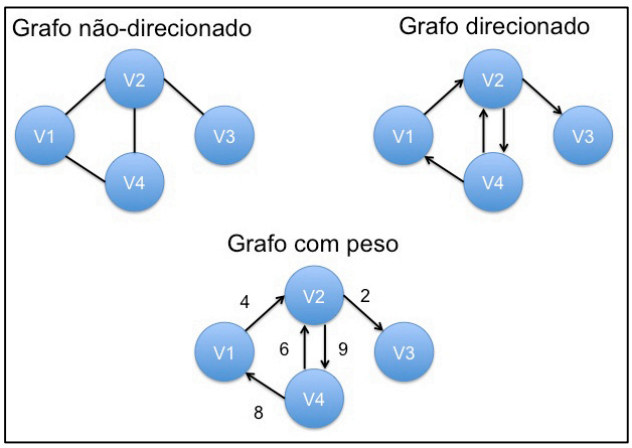


Figura 2 – Exemplo de grafos direcionado, não direcionado e com peso. Figura baseada no trabalho de Pavlopoulos et al. (2011).

3.2 Determinação sexual em vertebrados

A reprodução sexual é um processo biológico selecionado para a troca genética (Otto, 2009), gerando um novo espécime diferente de todos os outros pré-existentes. A troca/transferência de material genético é comumente presenciada tanto em eucariontes como em procariontes, sugerindo que o mecanismo de recombinação entre moléculas de ácidos nucléicos seja universal e bastante antigo (revisado por, Lodé 2012). Esse mecanismo de reprodução é bastante vantajoso no sentido de gerar variabilidade genética e de fato, um experimento demonstrou que as linhagens de bactérias sexuadas tiveram uma adaptação evolutiva mais rápida em ambientes rigorosos do que as linhagens assexuadas. Outro trabalho demonstrou que a taxa de mutação no crustáceo *water-flea* com reprodução sexuada é ¼ da taxa de mutação em indivíduos com reprodução partenogenética (revisado por, Ming e Moore, 2007). Em suma, os grupos de organismos com reprodução sexuada são compostos de indivíduos que trocam material genético por meio de células reprodutoras diferenciadas ou por transferência intercelular e, nesse contexto, pode-se compreender que esses organismos são de "sexos distintos".

Conceitualmente, a determinação do sexo nos vertebrados é definida como a decisão que conduz uma gônada bipotencial a se desenvolver como testículos ou ovários (Barske e Capel, 2008). Nos vertebrados, pode-se observar uma grande variedade de mecanismos de determinação do sexo, nos quais fatores genéticos e ambientais são fundamentais nesse processo. Gene(s)-chave, os quais podem ou não estar presentes em cromossomos sexuais, podem atuar como desencadeadores do início dessa cascata, no que é conhecido como uma determinação sexual mediada por fatores genéticos. Quanto aos fatores ambientais envolvidos na determinação sexual, é conhecido que a temperatura, pH e fatores sociais são importantes nessas vias em diversos grupos de vertebrados (revisado por, Manolakou et al., 2006; Sandra e Norma, 2010).

Em mamíferos therianos (inclui os marsupiais), o papel dos cromossomos sexuais (XY) na determinação do sexo é bastante conhecido. Nesse grupo, sabe-se que o cromossomo Y possui um gene chamado Sry (*Sex-determining region on the Y chromosome*), o qual codifica um fator de transcrição que inicia uma complexa cascata com ações sinérgicas e antagônicas que levam ao desenvolvimento de testículos nos machos (revisado por, Graves e Peichel 2010). Interessantemente, os monotremados (ornitorrinco e equidna) possuem 10 cromossomos sexuais (X_5Y_5), porém sem a presença do Sry em seu genoma (revisado por,

Graves e Peichel, 2010). Apesar da importância desse gene, o qual vem sendo extensivamente explorado em um grande número de trabalhos clínicos, evolutivos e funcionais, diversos outros genes são importantes para o processo de determinação sexual nos mamíferos.

No caso de aves é observado um sistema de cromossomos sexuais ZW. Classicamente, a hipótese de que uma proteína heterodimérica, produto da expressão de um gene com homólogos presente nos cromossomos ZW, possibilitaria o desenvolvimento de ovários; por outro lado, a ocorrência de uma proteína homodimérica, oriunda de um genótipo ZZ, levaria ao desenvolvimento de testículos (revisado por, Manolakou et al., 2006). Porém, um recente trabalho utilizando técnica de *knockdown*, demonstrou que o gene Dmrt1, o qual está presente somente no cromossomo Z, é um gene determinante de sexo em galinha que favorece o desenvolvimento de gônadas masculinas quando em dupla dose no genoma (genótipo ZZ) (Zhao et al., 2010). Em répteis, tais como os crocodilos, não tem sido reportado a presença de um sistema de cromossomos sexuais discerníveis, sendo a temperatura o principal fator na determinação do sexo. Em diversas espécies de tartarugas e lagartos é relatado a presença de sistemas XY, ZW e a influência da temperatura para a determinação do sexo (revisado por, Manolakou et al., 2006; Graves e Peichel, 2010).

Em anfíbios, tem sido relatada a presença de sistemas XY e ZW (revisado por, Graves e Peichel, 2010), embora a maioria das espécies não possuam cromossomos sexuais discerníveis (revisado por, Nakamura, 2010). A influência de cromossomos na determinação do sexo nesses grupos é ainda especulativa, tendo apenas um exemplo evidente relatado na literatura. Foi demonstrado que o gene DM-W (uma cópia truncada do Dmrt1), gene presente no cromossomo W da espécie *Xenopus laevis*, age como um gene determinante de sexo nessa espécie, favorecendo o desenvolvimento de ovários (Yoshimoto et al., 2008).

Os peixes possuem a maior diversidade e plasticidade de mecanismos de determinação do sexo, sendo relatada a existência de sistemas XY, ZW, sistemas com múltiplos cromossomos sexuais, sistemas X0 e Z0, influencia de genes autossômicos, influência ambiental, comportamental e reversão sexual (protrandria e protoginia, mediado por temperatura, pH ou por fatores sociais) (revisado por, Devlin e Nagahama, 2002; Mank et al., 2006; Sandra e Norma, 2010). O número de espécies de peixes com cromossomos sexuais heteromórficos relatados é muito baixo (Sandra e Normal, 2010), ocorrendo em torno de 10% das espécies segundo Arkhipchuk (1995), sendo que de todos os relatos, existe apenas duas evidências claras e diretas do papel desses cromossomos na determinação do sexo. Matsuda et

al. (2007) comprovaram que o gene Dmy (Dmrt1), presente no cromossomo Y do peixes medaka *Oryzias latipes*, possui um papel central na determinação do sexo nessa espécie, conduzindo ao desenvolvimento de testículos nos indivíduos portadores desse gene. O outro relato foi feito por Hattori et al. (2012), o qual demonstram o papel do gene amhy (similar ao Amh, hormônio Anti-Mulleriano) encontrado em indivíduos XY da espécie de peixe-rei *Odontesthes hatcheri*. Esse gene é expresso antes do gene Amh (conhecido por participar do desenvolvimento das células de Sertoli) e sua ausência conduz a expressão dos genes Foxl2 (*Forkhead box L2*) e Cyp19a1a (*Cytochrome P450 19A1a*), ambos conhecidos como participantes da determinação/desenvolvimento das gônadas femininas.

É interessante também que muitas das formas de determinação sexual nos peixes, bem como a origem dos cromossomos sexuais, surgiram de forma independente na evolução desse grupo (Mank et al., 2006; Graves e Peichel, 2010). A explicação vigente para essa grande diversidade de sistemas de determinação de sexo nos teleósteos é atribuída ao fato de o ancestral desse grupo ter sofrido eventos de duplicação gênica e genômica (Figura 3). Tais duplicações fornecem redundância para o genoma, abrindo potencial para que novos genes e vias metabólicas evoluam (revisado por, Mank et al., 2006; Venkatesh 2003). Dando suporte a essa conclusão, é bastante discutido que essas duplicações foram também responsáveis pela grande especiação e diversidade morfológica do grupo (Figura 4) (Venkatesh, 2003; Jaillon et al., 2009).

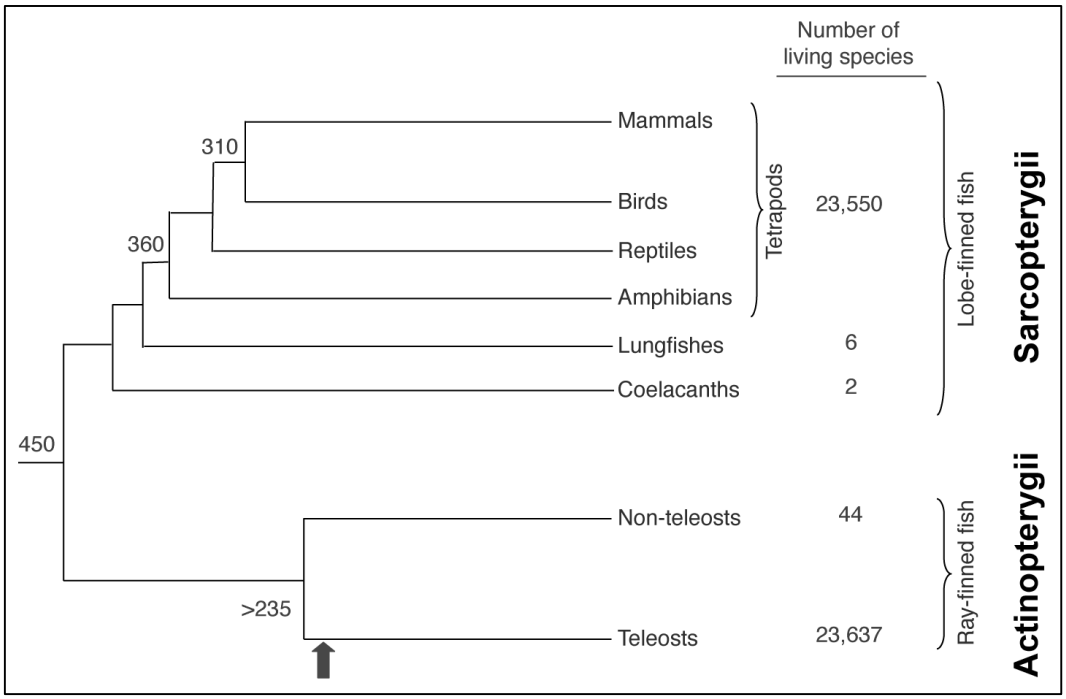


Figure 3 – Evolução dos Teleostomi, o qual enquadra os Actinopterygii e Sarcopterygii. Os números nos nós indicam o tempo de divergência em milhões de anos (Kumar e Hedges, 1998). A seta indica onde provavelmente ocorreu o evento de duplicação genômica total. Modificado de Venkatesh (2003).

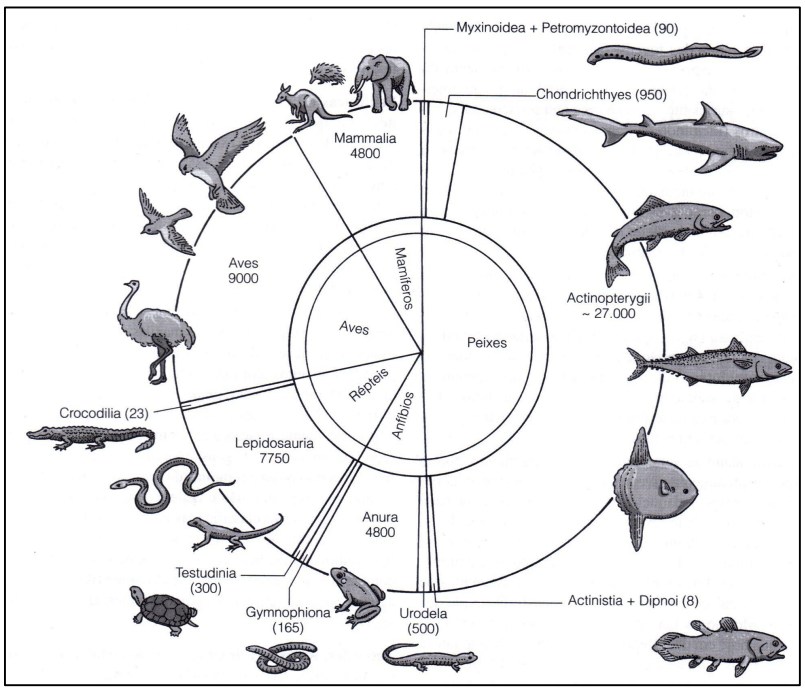


Figura 4 – Diversidade dos vertebrados. Os teleósteos compreendem aproximadamente 50% de todos os vertebrados (Pough et al., 2008).

Um grande número trabalhos vem relatando que diversos genes autossômicos são fundamentais para a determinação sexual nos vertebrados, sendo que o funcionamento e papel desse genes é relativamente conservado entre as espécies. Esses trabalhos envolvem análises de expressão gênica, silenciamento gênico, análises de promotores, hibridização *in situ* em tecidos e embriões, entre outros, e discutem tanto aspectos evolutivos como funcionais de diversos genes candidatos (revisado por, Morrish e Sinclair, 2002; Manolakou et al., 2006; Ijiri et al., 2008; Graves e Peichel, 2010).

Dentre os genes candidatos estudados, destacam-se o Sox9, Sox3, Foxl2, Dmrt1, Amh, Wt1, Wnt4, Fgf9, Cyp19, além de diversos parálogos comumente encontrados em peixes, os quais tem sido ranqueados como genes importantes nas vias de determinação e diferenciação sexual nos vertebrado como um todo (revisado por, Devlin e Nagahama, 2002; Morrish e Sinclair, 2002; Manolakou et al., 2006; Rodrigues-Marí et al., 2005; Ijiri et al., 2008; Sandra e Norma, 2010; Graves e Peichel, 2010). Apesar de diversos trabalhos abrangendo o assunto de determinação e diferenciação sexual terem sido publicados, tais mecanismos moleculares são ainda bastante obscuros, sendo melhor descritos em mamíferos do que em outros vertebrados. Dessa forma, há uma grande ausência de conhecimento sobre as interações entre os genes envolvidos nessas vias, ou seja, existe a suposição de inter-relação entre alguns genes porém, devido à grande complexidade e diversidade desse sistema, todos os trabalhos publicados até o presente momento estão muito aquém de relatar a maioria da interações que agem nessas vias, justificando o desenvolvimento de novos trabalhos na área.

A conclusão geral da determinação sexual nos vertebrados é que, exceto para escassos exemplos, a influência dos cromossomos sexuais na determinação do sexo em peixes, anfíbios e répteis, tem um carácter meramente especulativo, pois pouco se sabe do real papel desses cromossomos nesses grupos. O fluxo de informação que desencadeia o desenvolvimento das gônadas masculinas e femininas é pobremente conhecido e além disso, é bastante evidente que a determinação e diferenciação sexual em vertebrados, principalmente nos peixes, é algo bastante plástico e variável.

4. Objetivos e hipótese

4.1 Objetivos

Objetivo 1 - Geração de um método de predição de interação entre proteínas que possa ser aplicado em larga escala para qualquer organismo e proteína;

Objetivo 2 - Geração de redes de interação de proteínas para humanos e *Danio rerio* (*zebrafish*) para o estudo da evolução da determinação e diferenciação sexual em vertebrados.

4.2 Hipótese

É compreendido que muitos mecanismos relacionados ao desenvolvimento dos organismos possuem uma base extremamente conservada. Como exemplo, pode-se destacar os genes Hox, envolvidos com a determinação das porções dorsais, ventrais e laterais; ou a função do gene Pax6, gene envolvido no desenvolvimento da visão (ou foto-recepção). Além disso, Barske e Capel (2008) reportaram que muitos processos celulares, fatores de transcrição e vias de sinalização envolvidas na determinação do sexo e gonadogênese, são conservadas entre os vertebrados, sugerindo que a maquinaria básica pode ser muito similar. De fato, apesar da grande variedade de mecanismos de determinação sexual nos vertebrados, é observado uma grande conservação na determinação e desenvolvimento das gônadas tanto em nível histológico quanto molecular. Em nível molecular, é compreendido que o mesmo conjunto de genes atuem de forma semelhante nos diversos mecanismos existentes, podendo diferir principalmente no seu padrão temporal de expressão (revisado por, Graves e Peichel, 2010).

Nesse contexto, usando *zebrafish* e humano como modelos, o presente trabalho baseia-se na hipótese de que as interações proteicas dos genes que agem no início das vias de determinação sexual em vertebrados são muito conservadas, sendo essa conservação reduzida nos níveis finais desta via. A segunda premissa da hipótese, a qual trata de redução da conservação das interações, foi elaborada a partir do conhecimento de que o mecanismo de determinação e diferenciação sexual é bastante diversificado entre as espécies de vertebrados.

5. Capítulos

Os resultados da presente tese foram inseridos em formato de manuscritos nos capítulos 1 e 2.

5.1 Capítulo 1 – Desenvolvimento do UNISPPI

Nesse capítulo será relatado o desenvolvimento de um método de predição de interações entre proteínas denominado UNISPPI (*Universal In Silico Predictor of Protein-Protein Interactions*). Esse método foi desenvolvido devido ao fato de não existir na literatura um método adequado para predição dessas interações nos organismo-modelo adotados no presente trabalho.

Esse método é baseado em técnicas de aprendizado de máquina (em inglês, *machine learning*), o qual é uma excelente ferramenta para extração de padrões de um determinado conjunto de dados (Larrañaga et al., 2006). Como as informações contidas em um sistema biológico são de alta complexidade, o uso de aprendizado de máquina é de extrema utilidade para extração de informações desses dados (Larrañaga et al., 2006). De fato, essa técnica vem sendo cada vez mais utilizada na área de bioinformática, sendo empregada em análises de dados genômicos, proteômicos, *microarrays*, biologia de sistemas, evolução, mineração de texto, entre outros (Figura 5).

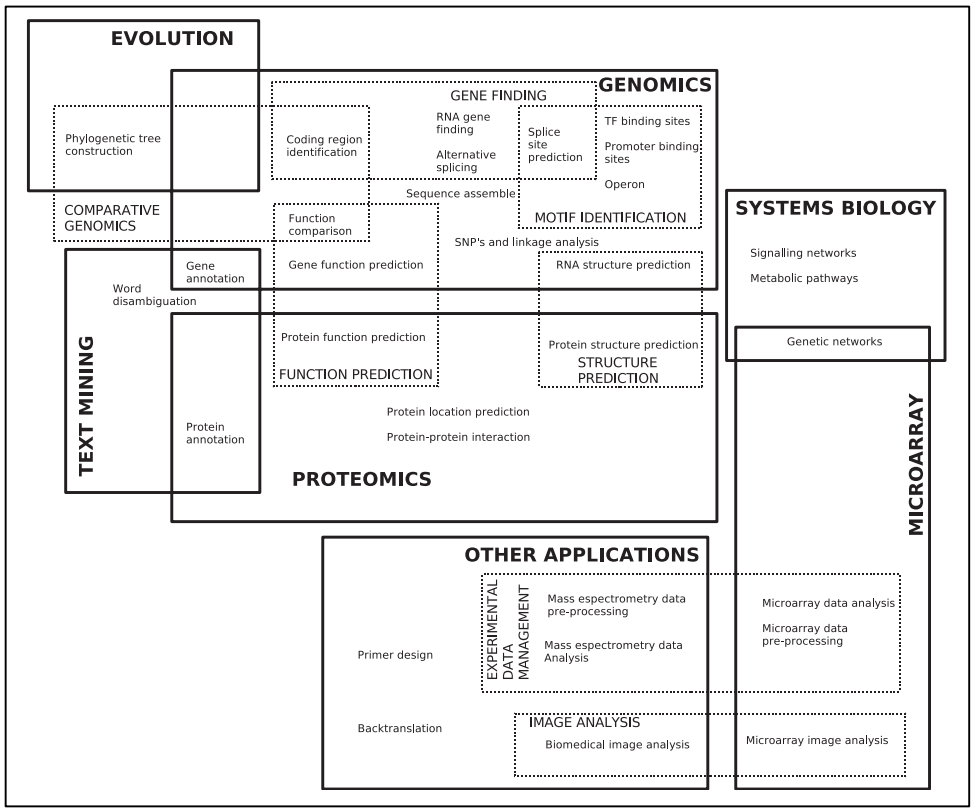


Figura 5 – Tópicos da bioinformática onde o aprendizado de máquina pode ser aplicado (Larrañaga et al., 2006).

Nesse caso, o aprendizado de máquina foi empregado em um problema de modelagem supervisionado (Larrañaga et al., 2006), o qual necessita de exemplos de proteínas que interagem e não interagem para o aprendizado (extração de padrões) e geração de um modelo de interação. Depois do processo de aprendizado, foram gerados modelos que foram aplicados na classificação de proteínas com interações desconhecidas para o algoritmo de aprendizado de máquina utilizado. Detalhes do procedimento metodológico, análises dos resultados e discussão estão inseridos no manuscrito que se segue.

O manuscrito foi submetido para a revista *Molecular System Biology* no dia 10/01/2013 e está formatado de acordo com as normas dessa revista. A numeração das figuras segue a ordem do manuscrito e não corresponde a ordem que vem sendo construída ao longo da tese. Todo o material suplementar deste manuscrito está inserido no CD que acompanha a tese.

The development of a universal *in silico* predictor of protein-protein interactions

*Guilherme Targino Valente[1], *Marcio Luis Acencio[2], Cesar Martins[1], Ney Lemke[2]

[1] Integrative Genomics Laboratory, Department of Morphology, Biosciences Institute, UNESP – Univ Estadual Paulista, Botucatu, Sao Paulo, Brazil. ZIP code: 18618-970.

[2] Laboratory of Bioinformatics and Computational Biophysics, Department of Physics and Biophysics, Institute of Biosciences, UNESP – Univ Estadual Paulista, Botucatu, Sao Paulo, Brazil. ZIP code: 18618-970.

*: Co-first authors.

Running Title: Predictor of protein-protein interactions.

Subject Categories: Bioinformatics; Computational Methods.

Total Characters Count Including Spaces: 26,158.

Abstract

Protein-protein interactions (PPIs) are essential for understanding the function of biological systems and have been characterized using a vast array of experimental techniques. These techniques detect only a small proportion of all PPIs and are labor intensive and time consuming. Therefore, the development of computational methods capable of predicting PPIs accelerates the pace of discovery of new interactions. This paper reports a machine learning-based prediction model, the Universal *In Silico* Predictor of Protein-Protein Interactions (UNISPPI), which is a decision tree model that can reliably predict PPIs for all species (including proteins from parasite-host associations) using only 20 combinations of amino acids frequencies from interacting and non-interacting proteins as learning features. UNISPPI was able to correctly classify 79.4% and 72.6% of experimentally supported interactions and non-interacting protein pairs, respectively, from an independent test set. Moreover, UNISPPI suggests that the frequencies of the amino acids asparagine, cysteine and isoleucine are important features for distinguishing between interacting and non-interacting protein pairs. We envisage that UNISPPI can be a useful tool for prioritizing interactions for experimental validation.

Keywords: amino acid frequencies/decision trees/machine learning/protein-protein interactions/universal predictor.

Introduction

Proteins are one of the most abundant classes of biomolecules that can interact with many other biomolecules in cells, such as DNA, RNA, metabolites and other proteins. The latter interactions – protein-protein interactions (PPIs) – are essential interactions that build functional units responsible for the functioning of all biological molecular pathways (see Bahadur, 2010, for a review). Consequently, building a list of all of an organism's PPIs can be useful for the molecular-level understanding of the core of complex traits. Therefore, the collection of all PPIs can be important for understanding the underlying mechanisms of diseases, facilitating the process of drug design, elucidating the functions of newly identified proteins, predicting their subcellular location and gaining insight into the evolution of some interaction or metabolic pathways, among other biological aspects of a cell or organism.

Physical PPIs can be identified using experimental methods, such as the yeast two-hybrid assay, mass spectrometry, protein microarrays, phage display, X-ray crystallography, fluorescence resonance energy transfer, surface plasmon resonance, atomic force microscopy and electron microscopy (Shoemaker & Panchenko, 2007a). However, these experiments, in addition to being expensive and time demanding, are not suitable for all proteins and do not report all interactions that can occur in cells or organisms (Han et al, 2005; Shoemaker & Panchenko, 2007a; Skrabanek et al, 2008). Therefore, a computational approach capable of reliably predicting PPIs can identify the potential interactions to be further interrogated using experimental approaches.

Several computational methods for predicting physical or functional protein interactions based on the information from several experimental and computational approaches, such as phylogenetic profile analysis, gene co-expression profiles, sequence co-evolution, synthetic lethality data, *in silico* two-hybrid systems, gene cluster and gene neighbor analysis, protein domain information (revised by Skrabanek et al, 2008; Shoemaker & Panchenko, 2007b), protein interface analysis, protein docking methods (Aytuna et al, 2005; Smith et al, 2002), and orthology and ontology information (De Bodt et al, 2009), among others, have been published. Moreover, several data sources can be used as examples for training several machine learning (ML) algorithms in methods of classification (Bock & Gough 2001; Chen & Liu, 2005; Hue et al, 2010; Shoemaker & Panchenko, 2007b; Zaki et al, 2009; Zhou et al, 2012; and therein). Additionally, it is possible to combine more than one approach and dataset to predict PPIs (Jansen et al, 2003; Zhang et al, 2004; Ben-Hur & Noble,

2005; Rhodes et al, 2005; Shoemaker & Panchenko, 2007b; and therein). However, some of the aforementioned methods use datasets that are not available for all species or all proteins. For instance, some methods use information from complete genome sequencing, gene-expression data or protein information that is not available for all sequences. Moreover, some methods have flaws: gene expression can not determine physical PPIs; molecular evolutionary events can disturb the prediction of PPIs in phylogenetic profile methods; functional relationships based on genomic organization are only suitable for prokaryotes; some methods are based on homology sequences, which is more difficult to determine than finding genes; and other computational and biological aspects may render the prediction very difficult or do not provide resolute results. Some of these points are argued in Chen & Jeong (2009), Pitre et al (2008) and Skrabanek et al (2008).

The primary protein structure, does not considering domains, motifs, ontology, phylogenetic trees and orthology, has sufficient information to estimate the propensity of PPIs (Anfinsen, 1973). Moreover, amino acid (aa) sequences are the most universal protein feature and thus appear to be ideal traits for use in building methods of predicting PPIs that are applicable to all proteins (Shen et al, 2007). In fact, many interesting and useful bioinformatics methods using primary sequences have been developed, and many methods include machine learning approaches (Bock & Gough, 2001; Guo et al, 2008, 2010; Liu et al, 2012; Lo et al, 2005; Martin et al, 2005; Nanni & Lumini, 2006; Ren et al, 2011; Res et al, 2005; Roy et al, 2009; Shen et al, 2007; Shi et al, 2010; Yu et al, 2010a, b; Zaki et al, 2009; Zhou et al, 2012). However, some of the methods based on ML and primary sequences have weaknesses, such as the building of negative datasets, the small number of examples and the large number of attributes (vectors) that code protein pairs.

To overcome these limitations, we present in this paper the Universal *In Silico* Predictor of Protein-Protein Interactions (UNISPPI), an ML-based approach that use features associated with amino acid sequences for building models for predicting PPIs. UNISSPI presents the following advantages over previously mentioned methods: (1) the negative training examples - the non-interacting protein pairs (hereafter named no-PPIs) - used to construct the training datasets are all experimentally supported no-PPIs; (2) the amount and diversity of PPIs and no-PPIs used as training examples are much higher than some of those used by other methods; (3) its computational cost is markedly reduced in comparison to other methods because of the small amount of learning features used to generate the models (only 20 features); and (4) the generated models are decision trees that can be easily analyzed, after

which the most important features capable of distinguishing PPIs from no-PPIs can be promptly identified. By only considering 20 attributes associated with amino acid sequences as learning features, UNISPPI can correctly classify 79.4% and 72.6% of known PPIs and no-PPIs, respectively, including protein pairs from eukaryotes, prokaryotes, viruses, and parasite-host associations. Because of these characteristics, we consider UNISPPI a truly universal predictor that is appropriate for large-scale prediction and thus can become a useful tool for indicating the most relevant interactions, which can be further validated experimentally.

Results

Predictive performance of generated models

The predictive performance of prediction models – in this case, decision tree models – generated from the Normal and Random training datasets are reported in the Figure 1 and Supplementary Figure S1, respectively. The predictive performances of the decision tree models generated from different groups of the Normal training datasets appear to be slightly different from each other (Figure 1), but it is evident that from their performance that models generated from the C or C' datasets resulted in the lowest predictive values. Interestingly, the Mann-Whitney U statistical test showed that the performance measures are significantly different for almost all of the performed comparisons (Table I). For these cases, we assumed that the best set of features that were able to discriminate the PPI and no-PPI classes were those that produced decision tree models with the highest median predictive values. Regarding the models with no significant differences, we assumed that both datasets used to generate the models had the same predictive performance. Thus, as seen in Table I, the models generated from F and F+C are the most predictive models, followed by those generated from F' and F'+C'. Moreover, the performance measures for models generated from these four training datasets vary by up to 1.4% (AUC), and the lowest variation is 0.6% (precision values for no-PPIs). Furthermore, the decision tree models created from Random datasets showed performance measures of approximately 0.5, confirming the hypothesis that the Normal datasets are able to extract relevant information about PPIs and no-PPIs (Supplementary Figure S1).

Analysis of decision tree models

Beyond their prediction capability, decision tree models can be used for knowledge acquisition to describe patterns in datasets. Decision trees are decision support tools inferred from the training data that use a graph of conditions and their possible consequences. The structure of a decision tree consists of a root node representing the most important condition for discriminating classes and internal nodes representing additional conditions for class discrimination under the main condition (Kingsford & Salzberg, 2008).

Of the 124 decision tree models generated by training the J48 algorithm on the F, F', F+C and F'+C' Normal training datasets (31 models for each dataset) (see "Materials and Methods" for details), we analyzed 24 decision trees (six for each dataset). The decision trees show that the combination frequencies of asparagine in the proteins from a protein pair (Asn-Asn) is the root node of the decision trees generated from all four datasets. Thus, this attribute can be considered the most important feature for discriminating the PPIs and no-PPIs of a dataset. Interestingly, some frequencies of Asn are not able to discriminate PPIs from no-PPIs (22 from all analyzed trees); in these cases, the combination frequencies of cysteines (Cys-Cys) or asparagine with cysteine (Asn-Cys) or asparagine with isoleucine (Asn-Ile) or isoleucines (Ile-Ile) in the proteins from a protein pair are relevant features for classifying the instances (Figure 2; Supplementary Table S1). Cysteine comprises 54.4% of the aas in the second level of the analyzed trees, whereas Ile and Asn comprise 26.5% and 19.1%, respectively, at the same level in all analyzed trees.

Classification of unlabeled instances

The combined models generated from the F and F' Normal datasets (see "Materials and Methods" for details) were used independently to classify the test set (Figure 3). In general, both the F and F' Normal combined model are able to correctly classify ~79.5% and ~73% of PPIs and no-PPIs, respectively (Table II). A more detailed analysis of the test set classified by the F' Normal combined model revealed that this set of features was able to classify species-specific PPIs and no-PPIs for prokaryotes, eukaryotes and viruses and PPIs and no-PPIs among different species (also including proteins from eukaryotes, prokaryotes and viruses). Some PPIs and no-PPIs from parasite-host associations were also correctly predicted (for instance, protein interactions between *Plasmodium falciparum* and humans;

between proteins of the human immunodeficiency virus, Epstein-Barr virus, influenza A virus, human cytomegalovirus, simian virus, chimpanzee hepatitis and human hosts; between the Sendai virus and mouse hosts; and between an enterobacteria phage and *Escherichia coli* host). Moreover, some no-PPIs among virus proteins and a non-host species were also correctly classified (Figure 4; Supplementary Table S2).

Moreover, the F and F' Normal training sets were described as unlabeled instances and were also classified using the F and F' Normal combined models, respectively. The results indicated that in general the instances were classified correctly (Supplementary Figure S2). Moreover, the F and F' Random combined models were applied to classify both the test set and the training set (in this case, described as unlabeled instances) separately, and the models were not able to classify those unlabeled instances (Supplementary Figure S3).

Discussion

In this paper, we presented a new method called a Universal *In Silico* Predictor of Protein-Protein Interactions (UNISPPI), based on primary sequence information and a small set of features, for classifying protein pairs as interacting or non-interacting proteins. To build the final UNISPPI, a large number of tests were conducted, and several protein physicochemical features were examined. The following discussion will show how and why we chose the core set of features for this method.

Selecting the best physiochemical features to form the UNISPPI and evaluating the universality of this method

Using an ML method, the ability of "frequency" and "composition" physicochemical features for prediction the PPI and no-PPI classes was assessed. The performance measures of decision tree models constructed from six Normal datasets (F, F', C, C', F+C and F'+C') showed that "composition" features do not provide the most relevant information for discriminating PPIs and no-PPIs in a dataset. However, the "frequency" features appear to have the most relevant information for predicting PPIs and no-PPIs. We are not arguing that the "composition" features are not important to PPIs; however, these features provide the most relevant information for discriminating a PPI based on our method. Regardless, the

statistical analysis indicated that the F and F+C datasets generated the models with the best predictive performances, followed by F' and F'+C'. However, the variation of the predictive values observed for these Normal datasets, in practical terms, do not differ expressively.

The use of decision trees was very helpful not just to extract information from the datasets but also to generate trees that yielded interesting results. The trees revealed that the combination of the frequencies of Asn, Cys and Ile were the most important feature for classifying the PPIs and no-PPIs. The frequency of Asn in symmetrical attributes (the frequency of Asn on both proteins) was the most important feature, even when the "composition" features were present. The Cys, Ile and Asn frequencies were present at the second level of the tree, in which symmetrical attributes containing Cys (the frequency of Cys on both proteins) were commonly reported. In conclusion, these trees lead us to conclude that the composition does not appear to convey relevant information for predicting PPIs and no-PPIs according to all analyzed trees, corroborating the conclusions obtained from analyzing ML predictive performances.

For the biological role of those three amino acids in PPIs, we considered several reports that showed the relationship between their frequency and PPIs. Analyzing the preference index of inter-chain contacts showed that the interactions between two Asn (Asn-Asn), two Cys (Cys-Cys), between Cys with other residues (including Ile), two Ile (Ile-Ile), and between Ile with other residues, have high preference index values (Cys-Cys presents the highest values) (for details, see Anashkina et al, 2007).

Regarding the interface features, there are interesting studies showing the participation of Asn, Cys and Ile in protein interfaces. Asn is one of the aa residues identified as having an affinity for interfaces (Jones & Thornton, 1996; revised in Archakov et al, 2003) and is present in the binding sites of antibody complementarity-determining regions (Padlan, 1990; Mian et al, 1991), for example. Moreover, Asn can play a role in stabilizing the protein structure of solvent-accessible regions of ubiquitin-associated domains, which might represent a binding surface (Mueler & Feigon, 2002). Ile and other hydrophobic and charged residues constitute 25% of the overall number of residues at protein-protein interfaces (Anashkina et al, 2007). Finally, Cys bridges are present in interfaces more frequently than expected (Ofran & Rost, 2003). Interestingly, Cys residues are able to make disulfide bonds, and these bonds are common inside or between two Cys residues and are important in cofactor binding, inter-

subunit interactions, DNA binding inhibition, membrane binding, subcellular localizations, stabilizing interactions and protein structures (Swaisgood, 2005).

Concerning the test of the models generated by ML training, we used the F and F' Normal combined models for classifying unlabeled instances of the F and F' test sets. These models were selected based on the differences in the predictive performances between F and F+C and between F' and F'+C', which were not statistically significant. Moreover, the "composition" features do not provide the most relevant information for predicting the interactions. Both models were equally able to classify the unlabeled instances with good predictive performance. These results were very exciting because these models were built using only eukaryotic examples, whereas the test set included a large number of prokaryote, eukaryote and virus examples. Additionally, the results showed that those models are in fact universal. Moreover, the ML training using Random datasets and the classification of the test set using the Random model provided support for all analyses and conclusions for this topic.

Finally, we concluded that using only the frequencies of amino acids in a protein pair was sufficient to classify PPIs and no-PPIs. According to the aforementioned facts, we decided that the best set of features for forming the UNISPPI is the F' Normal dataset. Although the F Normal dataset is slightly more predictive than the F' Normal dataset, both models have a similar ability to classify unlabeled protein pairs. Furthermore, the F' Normal dataset was chosen instead of the F'+C' Normal dataset because the differences between the predictive performances were not significant and because the decision trees showed that the "composition" information may be irrelevant. Moreover, the dataset chosen (F') to build the final UNISPPI contains only 20 attributes, which is more desirable than 841, 400 or 41 attributes. In large-scale projects, millions or billions of unlabeled protein pairs must be classified, and, in those cases, the F' dataset is obviously more useful than all other sets of features. These results support our conclusion that the primary sequence alone can be used to classify a PPI and showed that the predictive performance using only symmetrical attributes was sufficient.

The innovations of UNISPPI

To show the advantages of UNISPPI over other similar computational methods, we built a chart that showed the features of several similar computational methods (see method

descriptions and citations in Supplementary Table S3). Despite their usefulness, these methods show some limitations that we attempted to overcome through the development of UNISPPI. These limitations mainly encompass the construction of the negative example dataset and the number of instances and attributes in both training and test sets.

For the negative dataset in all other methods, the negative instances were obtained from predictions. These procedures are questionable because, it is easy to introduce bias into the dataset (for training and cross-validation procedures), which may be caused by the over-representation of negative examples and/or by intrinsic aspects of the negative dataset prediction procedure (for more detailed discussion see Park & Marcotte, 2011; Yu et al, 2010b). Moreover, the predictive performances differ according to the negative dataset prediction method used (Guo et al, 2008; Lo et al, 2005; Ren et al, 2011; Shi et al, 2010). Some methods rely on *in silico* artificial proteins from the PPI dataset to build the negative examples, and this procedure overestimates the results according to the level of randomness used during the artificial protein building. Moreover, it is intrinsically more difficult to distinguish real protein sequences than artificial ones because artificial sequences can lack specific protein patterns, such as motifs, domains or other signatures. Thus, the models generated using real protein sequences are more reliable and give better results for classifying real protein-protein interactions (Lo et al, 2005). In addition, some methods predict that if two proteins are not present as an interacting protein pair within the PPI dataset, those proteins can be included as a non-interacting protein pair to build the negative dataset. This statement does not make sense because it is not known whether these proteins do interact in an organism under some biological circumstances, such as pathogen attack, environmental influences, developmental stages and tissue-specific interactomes, among others (see method descriptions and citations in Supplementary Table S3).

Some methods are based on a low number of examples or are relevant to just one or few species. Thus, these methods can not be considered universal predictors because of their basis on a small universe of proteins (low diversity). In all methods with well-reported attribute constructions, the number of attributes is usually very high, which renders the methods unfeasible for large-scale prediction procedures. Moreover, some methods were not applied to classify instances that were not in the training set; thus, the true performance of these methods was not tested (see method descriptions and citation in Supplementary Table S3).

Direct comparisons among the predictive performance of the methods are not possible because of the differences in their features, interactions or methods for developing the negative dataset. Regardless, the UNISPPI has predictive values that are usually higher or comparable to those of the other methods (Supplementary Table S3).

Finally, UNISPPI is based on primary sequence information that is coded in 20 symmetrical attributes (F' construction), which is feasible for use in large-scale projects (approximately millions or billions of interactions). Moreover, because the UNISPPI model is based only on experimentally validated examples from diverse species and can be applied to predict PPIs or no-PPIs in a large number of species, the UNISPPI model is truly a universal predictor. This method can easily classify instances of a complete proteome or proteins from parasite-host associations when attempting to reduce the number of interactions to be further investigated using experimental procedures.

To learn how to apply UNISPPI in unlabeled instances, see Supplementary Text S2-S6. The F' Normal combined model (the UNISPPI per se) is available in the Supplementary File S1.

Materials and Methods

UNISSPI is a prediction model based on the combination of other models (see further details in the final of this section) constructed according to the typical steps performed in an ML procedure: 1 - collection of instances (in this case, known PPIs and no-PPIs) and selection of features to be used as learning attributes (in this case, a set of physicochemical features); 2 - construction of training datasets; 3 - selection of the ML algorithm to be trained (in this case, the J48 algorithm); 4 - generation and evaluation of the prediction model (Figure 5). The details of all procedures are presented in Supplementary Figures S4-S9. Moreover, important definitions are presented in Supplementary Text S1.

Collection and selection of instances for the training and test sets

The PPIs and no-PPIs were obtained from the BIOGRID (Stark et al, 2011) and Negatome databases (Smialowski et al, 2010), respectively and their aa sequences were retrieved from the UniProt database (UniProt Consortium, 2011) (details in Supplementary

Figure S4). From PPIs, we retrieved sequences from proteins associated with just one UniProt ID in BIOGRID. To avoid bias and misclassification, the PPIs and no-PPIs data were refined as follows: 1 – Only one example was chosen from redundant instances (example of redundancy: ab interaction = ba interaction, in which a and b represent proteins of one instance); and 2 - The instances classified as both PPIs and no-PPIs were also excluded.

Approximately 50% of the PPI examples for each species and all species-specific no-PPIs of eukaryotes were selected to build the training set. All other instances – instances from species with only one instance; instances with proteins from two different species; and instances from prokaryotes and viruses - were included in an independent test set (details in Supplementary Figure S4).

Generation of attributes, discretization procedure and dataset building

To build the attributes, we first calculated two different sets of physicochemical features for all proteins obtained: "frequency" and "composition". The "frequency" set holds the percentages for each of the 20 aa in the protein: $Af = \frac{NA}{N}$, where *NA* is the number of aa of type *A* (*NA*=1,2,3, …, 20), constituting a total of 20 feature descriptors. The "composition" set was obtained by grouping each aa of a protein into one of three different groups related to seven physicochemical features (details in Supplementary Table S4) followed by a calculation of the percentage of each group for each feature: $Cr = \frac{nr}{N}$, where "nr" is the number of the aa included into group 1, 2 or 3, constituting a total of 21 feature descriptors (3 groups X 7 physicochemical features) (details in Supplementary Figure S4). The "*N*" corresponds to the sequence length (Li et al, 2006). Proteins with identical features, as well as their instances in the training and test sets, were excluded from training and test datasets.

The discretization procedure was performed using the average and the standard deviation calculated for each feature descriptor for the "frequency" and "composition" sets based on the proteins in the training set (Supplementary Table S5). The values of each feature descriptor in the training set were binned into four groups according to the rules described in Table III. The protein pairs were then re-organized, and the feature descriptors of all instances that were already discretized were combined to form the final set of attributes. These procedures were conducted for all proteins for both the "frequency" and "composition"

training datasets separately, resulting in the F (from "frequency") and C (from "composition") datasets, respectively. The symmetrical attributes (attributes of the same physicochemical features) were selected from F and C, and two other datasets were constructed, the F' and C' datasets, respectively. Furthermore, F and F' were combined with C and C', respectively, forming the F+C and F'+C' datasets. The numbers of attributes for the F, F', C, C', F+C and F'+C' training datasets were 400 (20 X 20 features), 20 (400 - 380 features), 441 (21 X 21 features), 21 (441 - 420 features), 841 (400 + 441 features) and 41 (841 - 800 features), respectively. All training sets included the same instances (details in Supplementary Figure S5-S8).

The symmetrical attributes of all six training datasets were subjected to a standardization procedure because PPIs are supposed to be non-directional: the attributes with BA, CA, CB, DA, DB and DC were converted to AB, AC, BC, AD, BD and CD, respectively. This lack of directionality implies that f_1p_1 and f_1p_2 = f_1p_2 and f_1p_1 in a protein pair, where "f" and "p" mean "feature" and "protein", respectively (details in Supplementary Figure S9).

The "frequency" feature descriptors of all proteins from instances selected to form the test set were also subjected to the previously mentioned discretization procedures, and the average and standard deviation values obtained for the proteins of the "frequency" training set were used for the discretization (Supplementary Table S5). The instances were reorganized, and the procedures for selecting the symmetrical attributes were also conducted, resulting in the F and F' test datasets. Moreover, the symmetrical attributes of both test datasets were also subjected to the same standardization procedure as previously mentioned. All training sets included the same instances.

By the end, each training dataset (F, F', C, C', F+C and F'+C') consisted of 43,365 instances from 20 eukaryotic species and 12,510 different proteins (11,609 and 944 for PPIs and no-PPIs, respectively). Moreover, each test dataset (F and F') consisted of 42,473 instances from 78 species and 11,881 different proteins (11,576 and 305 for PPIs and no-PPIs, respectively) (Table IV-V; details in Supplementary Table S6-S7).

Balancing the training datasets and building the Random training dataset

Because the number of PPIs is much higher than no-PPIs and it is known that data imbalance degrades the performance of ML algorithms (Visa & Ralescu, 2005), it was

necessary to build balanced datasets from the six original training datasets. For this purpose, 31 balanced datasets were constructed, 30 containing 2,780 instances (all 1,390 no-PPIs and 1,390 PPIs) and 1 containing 550 instances (275 no-PPIs and 275 PPIs). Before the construction of these balanced datasets, the PPI examples were randomized in an attempt to avoid bias in the datasets, and the PPIs were randomly selected from the total set of PPIs for all balanced dataset constructions. All balanced datasets for all six training datasets were named Normal datasets. From these Normal datasets, we constructed Random datasets by shuffling their classes to check whether the generated model on non-shuffled datasets learned the traits actually associated with PPIs instead of traits associated with any random subset of protein pairs. Thus, we generated 62 datasets for each of the six balanced training datasets, half of which were Normal and half were Random datasets (details in Supplementary Figure S9).

Predictive performance: validation and application tests

The prediction models were generated by training ML algorithm J48 (Quinlan, 1993) with bootstrap aggregating (bagging) (Breiman, 1996) on all balanced datasets (Normal and Random datasets) for each group of the six training datasets. Before constructing the UNISPPI per se (a model based on the combination of 31 F' Normal models), all generated models were first validated by estimating their predictive performances using the 10-fold cross-validation test via WEKA (Waikato Environment for Knowledge Analysis) software version 3.7.5 (Witten et al, 2011) (details in Supplementary Figure S9).

To test the ability of UNISPPI to classify unseen instances, we performed an application test. Four combined models were obtained from the 31 models generated by the ML training using the F and F' Normal and Random training datasets separately. These combined models were then used to classify the F and F' test datasets (Table V, Supplementary Table S6-S7; procedure detailed in Supplementary Text S2- S6). Comparing the ability to classify unseen instances using models from normal and random datasets is useful for verifying the predictive ability. The same procedures were performed for the instances used in the F and F' training datasets (Table IV; Supplementary Table S6-S7). No instances in the test set were previously exposed to the J48 algorithm.

Performance measures and statistical analysis

The estimated predictive performance of the prediction models was measured in terms of their precision, recall (equivalent to sensitivity or true positive rate) and area under the receiver operating characteristic curve (AUC). The performance measures estimated for models constructed from Normal datasets were compared using the non-parametric Mann-Whitney U statistical test (Wilcoxon, 1947). In the application test, instances with a classification score of 0.50 were not classified neither PPIs nor no-PPIs.

Acknowledgements

We thank the FAPESP (São Paulo Research Foundation) and CNPq (The National Council for Scientific and Technological Development) for financial support.

Author Contributions

GTV and MLA conceived the research proposal and designed the experiments. GTV performed the data acquisition. NL supervised the experimental design and the machine learning procedures. GTV, NL, MLA and CM analyzed and interpreted the results and wrote and revised the manuscript.

Conflict of Interest

The authors do not have any conflicts of interest.

References

As referências deste manuscrito estão listadas no tópico “7. Referências bibliográficas”.

Figure Legends

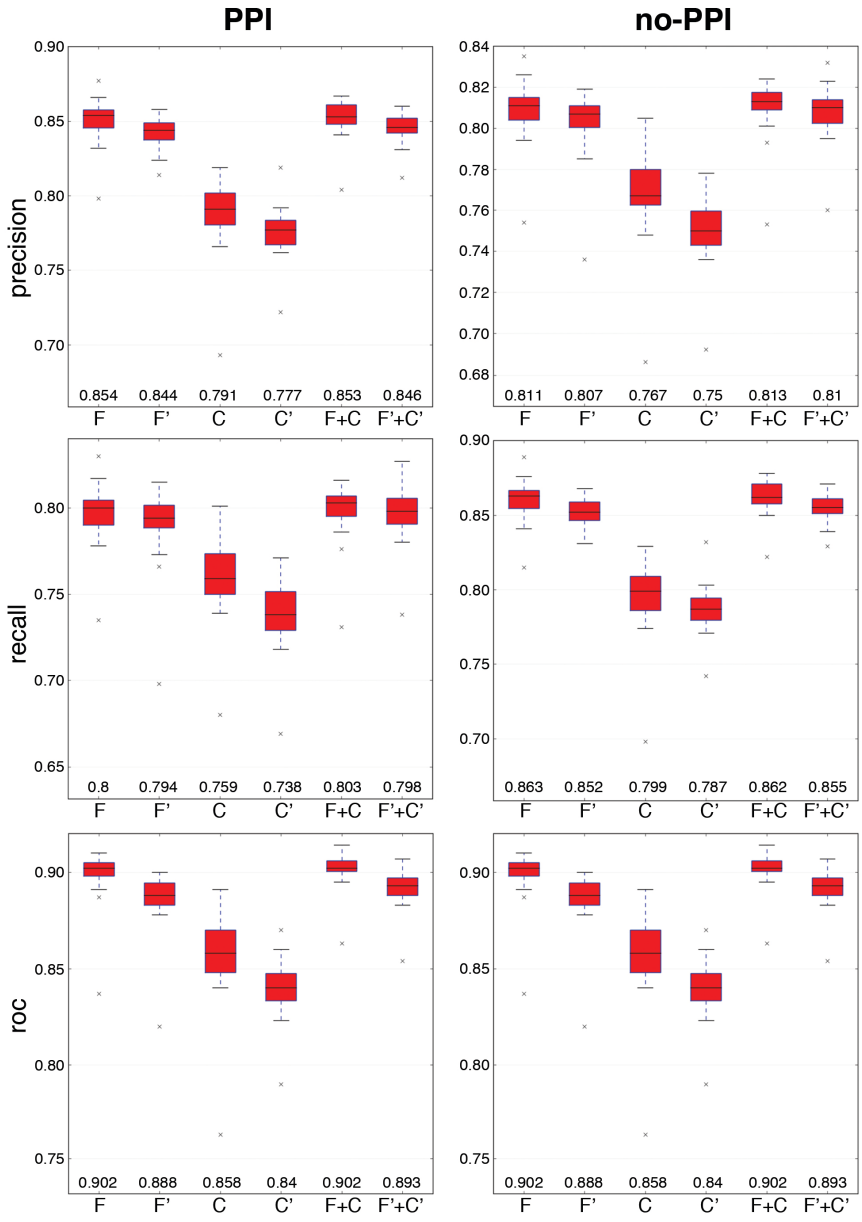


Figure 1 – Predictive performance of the machine learning using the Normal training datasets. The letters F, C and F+C indicate that the Normal training datasets originated from the feature descriptors "frequency", "composition" and "frequency" plus "composition", respectively. The prime symbol indicates the Normal training datasets formed using the symmetrical attributes of the previously mentioned datasets (details in the "Material and Methods" section). The numbers at the bottom of the boxes are the medians for each dataset.

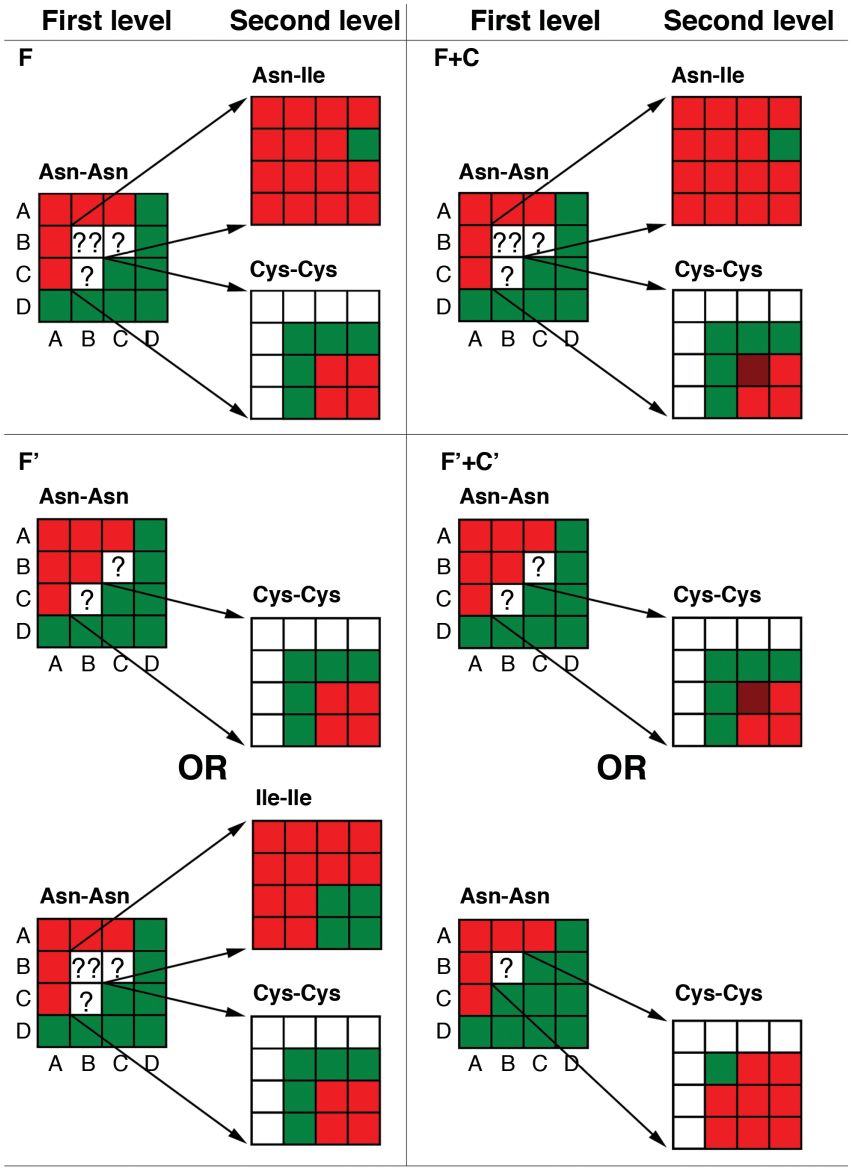


Figure 2 – Synthesis of several decision trees generated during the ML training. The aa at the top of each box is relative to the attributes specified at the first and second level of the trees. A, B, C and D indicate the low, moderate, high and very high bins, respectively (for more details, see Table III and the "Materials and Methods" section). Green boxes, a combination that classifies an instance as a PPI; Red boxes, a combination that classifies an instance as a no-PPI; Brown boxes, a combination that classifies an instance as a PPI or no-PPI; "?", a combination for which an instance can not be classified, requiring classification at the next level of the tree.

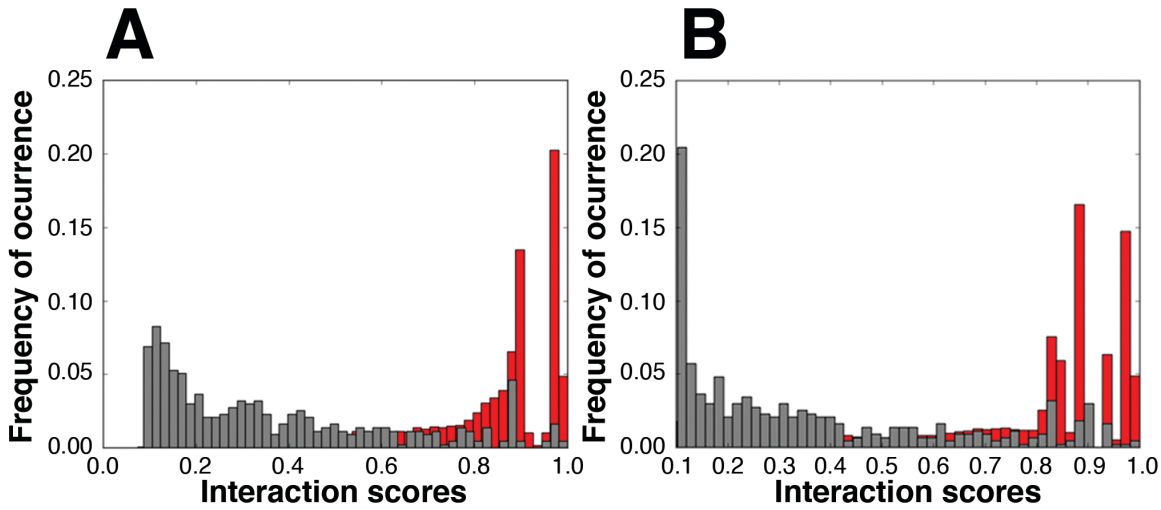
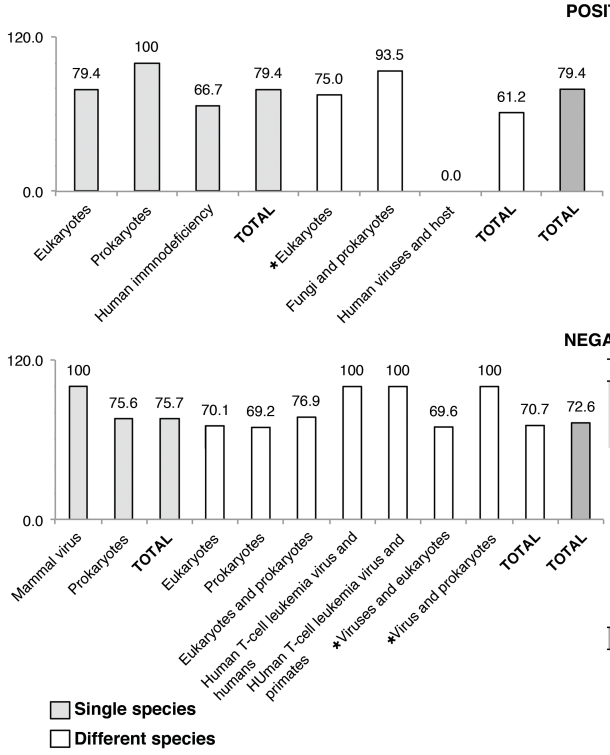


Figure 3 – Graphics of classification of the test sets using the F and F' Normal combined models. The score represents the classification of the instances as PPI. The instances were classified as PPIs or no-PPIs, and no-PPIs classification scores were converted to interaction scores. "A", classification of the F test set using the F Normal combined model; "B", classification of the F' test set using the F' Normal combined model; Red, PPIs instances; Gray, no-PPIs instances.



Groups	TOTAL	PPIs
Eukaryotes	41967	33319
Prokaryotes	1	1
Human immunodeficiency virus	3	2
TOTAL	**41971**	**33322**
Eukaryotes	16	12
Fungi and prokaryotes	31	29
Human viruses and host	20	0
TOTAL	**67**	**41**
TOTAL	**42038**	**33363**

Groups	TOTAL	NOPPIs
Mammal virus	1	1
Prokaryotes	168	127
TOTAL	**169**	**128**
Eukaryotes	177	124
Prokaryotes	26	18
Eukaryotes and prokaryotes	13	10
Human T-cell leukemia virus and humans	2	2
Human T-cell leukemia virus and primates	1	1
Viruses and eukaryotes	46	32
Virus and prokaryotes	1	1
TOTAL	**266**	**188**
TOTAL	**435**	**316**

Figure 4 – Detailed analysis of the classification of the test set using the F' Normal combined model. Single species indicates PPIs and no-PPIs within the same species; Different species indicates PPIs and no-PPIs among different species; *, groups that include instances of parasite-host associations.

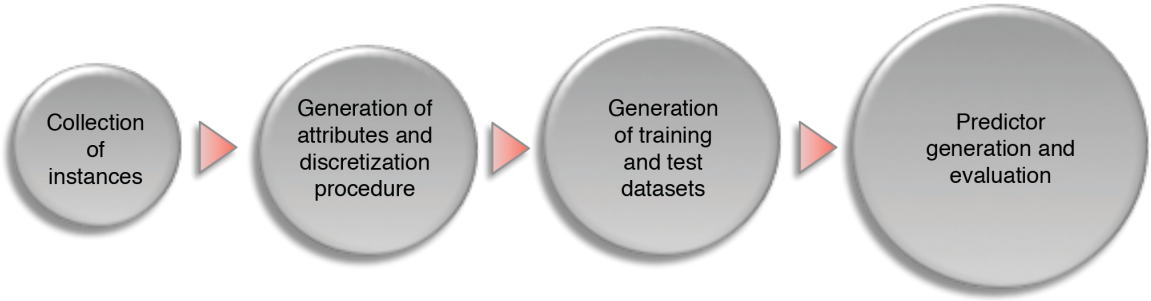


Figure 5 – A general workflow of the procedures adopted in this work.

Table I – Mann-Whitney U statistical test applied to performance measures of models generated from Normal training datasets. *, no significant difference.

		PPI	NO-PPI
Datasets		**p-value**	**p-value**
CxC'	Precision	1.15E-04	2.43E-06
	Recall	3.51E-06	0.002
	AUC	1.73E-06	1.73E-06
CxF+C	Precision	1.18E-11	1.61E-10
	Recall	5.36E-10	8.36E-12
	AUC	2.26E-11	2.26E-11
CxF'	Precision	8.41E-12	5.61E-10
	Recall	4.38E-09	0.000
	AUC	1.51E-09	1.51E-09
CxF'+C'	Precision	8.36E-12	1.61E-10
	Recall	9.90E-10	7.21E-12
	AUC	1.17E-10	1.17E-10
C'xF+C	Precision	7.54E-12	2.17E-11
	Recall	5.08E-11	0.000
	AUC	7.46E-12	7.46E-12
C'xF'+C'	Precision	7.49E-12	1.49E-11
	Recall	2.76E-11	7.56E-12
	AUC	8.33E-12	8.33E-12
F+CxF'+C'	Precision	9.00E-04	1.76E-02

	Recall	9.15E-02	*	0.002	
	AUC	2.46E-08		2.46E-08	
FxC	Precision	1.72E-11		2.11E-10	
	Recall	7.28E-10		1.11E-11	
	AUC	1.40E-10		1.40E-10	
FxC'	Precision	7.49E-12		2.18E-11	
	Recall	3.84E-11		0.000	
	AUC	4.59E-11		4.59E-11	
FxF+C	Precision	2.49E-01	*	8.91E-02	*
	Recall	1.27E-01	*	2.79E-01	*
	AUC	1.34E-01	*	1.34E-01	*
FxF'	Precision	6.91E-04		1.61E-02	
	Recall	7.23E-02	*	0.001	
	AUC	4.62E-08		4.62E-08	
FxF'+C'	Precision	1.25E-02		1.97E-01	*
	Recall	4.14E-01	*	1.61E-02	
	AUC	4.67E-06		4.67E-06	
F'xC'	Precision	7.54E-12		1.11E-10	
	Recall	1.40E-10		0.000	
	AUC	1.16E-10		1.16E-10	
F'xF+C	Precision	3.14E-05		2.31E-04	
	Recall	2.86E-03		1.50E-04	

	AUC	4.57E-10		4.57E-10	
F'xF'+C'	Precision	9.50E-02	*	9.49E-02	*
	Recall	9.75E-02	*	0.162	*
	AUC	1.12E-02		1.12E-02	

Table II – Summary of the classification of the test set using the F and F' Normal combined models. numb., number.

	Classification using F Normal combined model		**Classification using F' Normal combined model**	
Classes	**numb. of instances**	**% of instances**	**numb. of instances**	**% of instances**
PPI	33,609	79.9	33,363	79.4
no-PPI	319	73.3	316	72.6

Table III – Descriptions of the bins used to discretize the attributes of the F and C datasets. avg, average; stdv, standard deviation.

Bins	**Description**	**Letter codes**
Low	< avg – stdv	A
Moderate	> avg – stdv and < = avg	B
High	> avg and < = avg + stdv	C
Very high	> avg + stdv	D

Table IV – Descriptions of instances inserted in the training set.

	Number of instances		
Groups	**PPI**	**no-PPI**	**Number of species per group**
Amphibians	0	1	1
Annelids	0	6	3
Arthropods	6,295	1	1
Avians	9	3	1
Mammals	499	1,320	5
Plants	768	37	6
Rays	0	4	2
Yeasts	34,404	18	2
TOTAL	**41,975**	**1,390**	**-**

Table V – Descriptions of instances inserted in the test set.

	Number of instances		
Groups	**PPI**	**no-PPI**	**Number of species per group**
	Single species		
Eukaryotes	41,967	0	14
Prokaryotes	1	168	17
Virus	3	1	2
TOTAL	**41,971**	**169**	
	Different species		
Eukaryote viruses and eukaryote non-host	0	13	14
Eukaryote parasites and host	21	37	17
Eukaryotes	15	176	34
Eukaryotes and prokaryotes	31	13	16
Prokaryote viruses and host	0	1	2
Prokaryotes	0	26	8
TOTAL	**67**	**266**	**-**
TOTAL SINGLE AND DIFFERENT SPECIES GROUPS	**42,038**	**435**	**-**

Supplementary Information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb/).

5.2 Capítulo 2 – Análise da determinação e diferenciação sexual em vertebrados utilizando redes de interação entre proteínas

Nesse capítulo, o método de predição de interação entre proteínas UNISPPI foi aplicado para todas as proteínas de *zebrafish* e humanos, gerando grafos para todo o proteoma dessas espécies. Em seguida foi avaliado a evolução dessas redes bem como aspectos peculiares de uma sub-rede envolvida nas vias de determinação e diferenciação sexual em vertebrados. Segue abaixo o *draft* inicial no formato de manuscrito (ainda não submetido à publicação) com os resultados obtidos no presente trabalho.

A numeração das figuras e tabelas segue a ordem do capítulo 2 e não corresponde a ordem que vem sendo construída ao longo da tese.

The zebrafish and human protein-protein interaction network: evolutionary aspects of sex determination and differentiation

*Guilherme Targino Valente[1], Marcio Luis Acencio[2], Ney Lemke[2], Cesar Martins[1]

[1] Integrative Genomics Laboratory, Department of Morphology, Biosciences Institute, UNESP – Univ Estadual Paulista, Botucatu, Sao Paulo, Brazil. ZIP code: 18618-970.

[2] Laboratory of Bioinformatics and Computational Biophysics, Department of Physics and Biophysics, Institute of Biosciences, UNESP – Univ Estadual Paulista, Botucatu, Sao Paulo, Brazil. ZIP code: 18618-970.

Running title: Evolution of PPIs network for sex determination

Subject categories: Bioinformatics.

Correspondence author: Guilherme Targino Valente

Abstract

The sex reproduction is an important mechanism to increase the genomic variability into the species. The most understood mechanism of sex determination and differentiation (SDD) in vertebrates is related to therian mammals, which Sry gene triggers the male development. The sex determination in fishes is much more diverse, in which the presence of a key gene in this process is poorly reported; thus, the sex determination seems to be trigged by several different factors. The present work used a computational predictor of protein-protein interactions to reconstruct the whole protein network for zebrafish and humans. The results reported a relative low conservation among those networks, including the sub-networks of SDD. On the other hand, we discuss that despite of this low conservation, the networks conserved the basic aspects involved in the SDD. Moreover, we showed that FOXO5, CYP19A1B of zebrafish and FOXO3 and CYP19A1 of humans worked like hubs within their sub-networks. It was also reported a conservative role of SOX proteins (SOX9A, SOX3 of zebrafish and SRY of humans) and probably SOX3 is much more important to the sex determination of non-therian vertebrates than previously reported. In conclusion, evolutionary aspects of vertebrate protein interactions and vertebrate sex determination are under the focus of this work and new candidates were reported to further experimental procedures.

Key words – Vertebrates, evolutionary biology, machine learning, protein network

Introduction

The vertebrate sex determination and differentiation (SDD) mechanisms are very diverse, in which key genes and complementary genes (including hormones and transcription factors), sex chromosomes, social cues and environment can act along this pathways in attempt to develop male or female gonads from bipotential gonads (Manolakou et al., 2006; Sandra and Norma, 2010). Despite of this theme have being exhaustively exploited by several researchers using different approaches, it is poorly understood how this pathway flows to achieve the final aspect of male and female development.

In mammals (except platypus and echidna), it is well known that the Sry (Sex-determining region on the Y chromosome) gene, present in the Y chromosome, codes a transcription factor that triggers the male gonad development (Graves and Peichel, 2010). On the other hand, the presence of genes analogous to the Sry in other vertebrates is scarce, being known just three examples: 1 – the Dmy gene (paralog to the Dmrt1, doublesex and mab-3 related transcription factor 1) is found in the Y chromosome of the fish *Oryzias latipes*, starting the sex male determination (Matsuda et al., 2007); 2 - the DM-W (paralog to the Dmrt1, doublesex and mab-3 related transcription factor 1) is found in the W chromosome of the South African clawed frog *Xenopus laevis*, which is the sex determining gene prone to female development in this species (Yoshimoto et al., 2008); 3 – it was recently found a gene paralog to the Amh (Anti-Mullerian hormone), named Amhy, which is found in XY individual of the Patagonian fish pejerrey (*Odontesthes hatcheri*) when is knocked-down, conduct to the Foxl2 (Forkhead box L2) and Cyp19a1a (Cytochrome P450 19A1a) overexpression, suggesting that this gene may be a master sex determining gene in this species (Hattori et al., 2012).

Moreover, several other genes work conservatively to the sex determination across the vertebrates, such as the Sox9, Amh, Wt1, Sf1, Dax1, Dmrt1, Fgf9, Foxl2, Cyp19, among others (Morrish and Sinclair, 2002; Manolakou et al., 2006; Rodrigues-Marí et al., 2005; Ijiri et al., 2008; Graves and Peichel, 2010; Sandra and Norma, 2010; Zhao et al., 2010). However, the flow of information across these genes and proteins to develop gonads from bipotential gonads to latter male or female stages is poorly understood, even to mammal models such as mouse and humans. Nowadays, the flow of information to develop the phenotypic aspects of a cell or organism have been accessed using complex network theory, using several biological information such as promoter sequences, metabolic information, microRNAs, proteins,

among other (Junker, 2008). In this way, the present study focus in the analysis of vertebrate SDD pathways, using the protein-protein interactions (PPIs) networks of the human and zebrafish proteomes as models. This paper reported a general low conservation between the networks, including the sub-networks related to SDD but reported that a conservative role of SOXs proteins and elected FOXO5/FOXO3 as new candidates to be further experimentally checked.

Material and methods

Proteome set and PPI predictions

The *Homo sapiens* and *Danio rerio* Reference Proteome Sets were downloaded from UniProt database (UniProt Consortium, 2011), constituting of 70,108 and 41,405 proteins, respectively. All protein pair possibilities were built for both species, considering as redundancy the interaction A with B and B with A; just one of those combination from all combinations were inserted in the dataset, constituting two non-redundant set of unlabeled instances datasets (protein associations to be classified as PPIs or non-PPIs).

The set of attributes (properties of instances) used here were the amino acids frequencies and they were obtained and organized according the UNISPPI (Universal *In Silico* Predictor of Protein-Protein Interactions) protocol (Valente et al., in preparation). UNISPPI is a computational method, based on machine-learning algorithms, able to predict PPIs for all species and proteins using only a small set of amino acid frequency combinations.

A total of 842,366,535 and 2,457,600,885 instances, from zebrafish and human, respectively, were submitted to UNISPPI classification according the recommended protocol (Valente et al., in preparation). The cutoff applied to recover feasible PPIs was based on the score average of the instances wrongly classified after the UNISPPI application test (a test to evaluate the power of UNISPPI to classify unlabeled instances). Thus, the PPIs scores > 0.728 was chose as true positive results.

Orthologs search and sex determination and differentiation sub-network extraction

To compare the human and zebrafish PPI networks, it was necessary to find the ortholog proteins between both species. It was performed the blast search of all zebrafish proteins against all human proteins and *vice-versa*, using the parameters (blastp –query IMPUT.fasta –db DATABASE.fasta –outfmt 10 –best_hit_overhang 0.4 –best_hit_score_edge 0.1 –use_sw_tback –out OUTPUT) and the best hits for both blast searches was the E value cuttof < 0.72, that is an average of all E values of both blast searches. The reciprocal best-hits were collected using the Python script called find-reciprocal.py (http://ged.msu.edu/angus/tutorials/reciprocal-blast.html) with a E value cuttof ranging from 0.03 to 0.04 (constitute the average of E value for each reciprocal best-hit procedure). Each PPI that had both proteins with orthologs and the interologs (conserved interactions) were selected to evaluate the evolution of those PPI networks.

Based on the literature evaluation, the SDD candidate proteins for zebrafish and humans were collected constituting a large sub-network (Table 1). Some PPIs among those proteins were selected constituting another sub-networks called D-SDD and H-SDD (*Danio* and *Human* Sex Determination and Differentiation, respectively). All sub-networks were evaluated according to their degree of distribution (number of connections) (Steuer and López, 2008) and specific comparisons were performed using the Cytoscape 2.8 (Smoot et al., 2011). The shortest paths were obtained for some proteins of D-SDD and H-SDD.

Results

For zebrafish and humans, 103,260,362 and 266,786,619 instances were classified as true positive PPIs, respectively. The blast searches performed between both proteomes, reported that ~97% and ~75% of *D. rerio* and *H. sapiens* proteins, respectively, blasted each other. The reciprocal best-hit analysis reported 22,413 orthologs between both proteomes, that correspond to ~56% and ~39% of blasted proteins for zebrafish and humans, respectively. Based on the orthologs list, it was possible to recover that 24% and 8.3% of zebrafish and humans interactions have both proteins with an ortholog, respectively (Figure 1, 2).

It was possible to determine that 7,211,150 interologs, considering one interaction as conserved if their proteins interact in both species. This value represents < 7% of the global network (entire network) of both species, but represents 29% and 32.5%, of zebrafish and

humans, respectively, considering the networks which all proteins have an ortholog (Figure 2). Moreover, the degree of distribution (number of connections) reported very divergent values for each ortholog protein for all SDD sub-networks (Table 2, 3).

At least five and four hubs (highly connected nodes) can be reported for D-SDD and H-SDD, respectively (red and orange nodes), but just two of them were conserved hubs (FOXO5/FOXO3 and CYP19A1B/CYP19A1). Moreover, apparently just the FOXO5/FOXO3 have highly amount of interologs (Figure 3). The shortest paths were retrieved ranging from three SOX proteins (SOX9A and SOX3 for zebrafish and SRY for humans) to CTNNBIP1 and it was reported that the conserved hubs aforementioned worked as intermediate conserved nodes into those three small networks (Figure 4). On the other hand, analyzing the interolog percentages for D-SDD and H-SDD, it is possible to see very divergent values for each conserved node between those sub-networks (Figure 5).

Discussion

General conservation between the networks

The global PPI networks for zebrafish and humans have a small level of conservation that we suppose to occur because the low amount of orthologs between both species. However, if the conservation of the networks which both proteins have an ortholog to be considered, these levels increased from < 7% up to 32.5%. Moreover, the vertebrate SDD sub-networks seem to be also relatively low conserved, if considering the degree of distribution. Considering all four SDD sub-networks (two for zebrafish and two for humans), we suggest that despite of homologous proteins are present, these networks evolved very divergently but maintained the same SDD pathway.

We suppose that several factors could have conducted to the divergence between both networks here observed by adding or deleting proteins, modifying ortholog proteins across the evolution, or rewriting the edges across the time (that do not implicate necessarily the gain or loss of new genes, but are related to gain or loss of edges) (revised by Chae et al., 2012). All those factors can insert or delete the edges across the evolution and taken together, these aspects could have influenced the general and hub conservations (at least to the FOXO5/FOXO3 interactions in D-SDD and H-SDD) here observed and could give the differential edges reported for both small SDDs. The previous statements are based taking

into account several evolutionary mechanisms: 1) intrinsic aspects of fish genome evolution, that is characterized by the arisen of paralogs followed by neo or sub-functionalization process occurred after the whole genome duplication (WGD) in the ancestral of teleost fishes (Venkatesh, 2003; Taylor and Raes, 2004); 2) protein regions are under divergent evolutionary constraints (Bloom et al., 2006; Dunker et al., 2002) and highly expressed proteins evolve slowly (Drummond et al., 2005); 3) evolutionary aspects of the networks, which hub proteins have higher conservation level (positive correlation between the average degree and the interaction conservation) (Fox et al., 2009), evolve slowly (Chae et al., 2012), and protein complexes interactions are more conserved (revised by Zinman et al., 2011).

The reasons of low conservation of the networks have been not extensively discussed (revised Zinman et al., 2011), but there is no consensus about the evolutionary rate of protein networks (revised by Börnke, 2008; Zinman et al., 2011). Zinman et al. (2011) argued that this disagreement occurs because it is usually evaluated only small datasets, interaction data and specific conditions. The authors compared networks among four eukaryotes and found that although individual interactions are low conserved, the interactions within the modules are higher conserved. A general conservation at the modules < 20% and the top conservation at ~40% (the last one between *Caenorhabditis elegans* and *Saccharomyces cerevisae*) was reported by Zinman et al. (2011). However, the conservation rates between the entire PPI networks of two closely related species (*S. cerevisae* and *S. pombe*) vary from 1.78 and 57.96 (Zinman et al., 2011). Another work reported that < 1% of the instances overlap among the protein interactions for humans, *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster* networks (Gandhi et al., 2006). The conservation of global networks here reported is correlated to the level of PPI network conservation aforementioned.

We conclude that the general conservation found in this work is related to the molecular and genomic evolutionary aspects. Moreover, we included thousand or million of interactions from uncharacterized proteins and proteins without orthologs, that certainly could decrease the conservation between both species.

The protein networks of sex determination and differentiation

The proteins FOXO5/FOXO3 are hubs within D-SSD and H-SDD. It was reported that Foxo5 is highly overexpressed in absence of Eaf1 or Eaf2 (ELL associated factor), as

well as the Wnt4a. The Eaf1 and Eaf2 is known to be highly expressed in several tissues including brain, placenta and testis in humans (Simone et al., 2001; Xiao et al., 2003) and the Wnt4 is clearly observed to act in the gonad development in mammals (Bernard and Harley, 2007). Despite that Foxo5 to be not directly observed as an important gene in the sex determination, its gene ontology (GO 0001556) showed that it is important to the oocyte maturation. FOXO5/FOXO3 have several interologs including connections with proteins previously known to work in the sex determination and differentiation and brain development pathways. The interologs observed for FOXO5/FOXO3 include the edges with ATRX, CYP19A1B/CYP19A1, FGF20/FGF9, CTNNBIP1, SOX9A/SOX9 and WNT4A/WNT4. The interactions with ATRX and SOX9A/SOX9, probably represent antagonistic interactions while the interactions with CYP19A1B/CYP19A1 could be synergic interactions. The ATRX is a protein important to the development of Leydig cells, which are adjacent to the seminiferous tubes at testis (Wilhelm and Koopman, 2006); SOX9a/SOX9 are transcriptor factors that works in the beginning of male development and is considered a key gene in the sex determination for non-therian vertebrates (revised Rodrigues-Marí et al., 2005; Manolakou et al., 2006; Ijiri et al., 2008); the Cyp19 class codes aromatases that convert androgen into estrogen and some of them are known to work in the initial stages of sex determination in several species (Morrish and Sinclair, 2002; Ijiri et al., 2008). Moreover, despite of CYP19A1B of zebrafish is known as brain aromatase, it is suggested also to work in sex differentiation (Trant et al., 2001), while the CYP19A1 works in the sex determination in several species (Morrish and Sinclair, 2002; Rodriguez-Marí et al., 2005; Ijiri et al., 2008).

No conclusions can be done to the edges between FOXO5/FOXO3 with WNT4A/WNT4, because WNT4 protein works both in male and female development and in several embryogenesis processes of mammals (Bernard and Harley, 2007). The same can be observed for CTNNBIP1, which in mouse is responsible for the inhibition of interaction of Beta-catenin with other proteins, working at the cell proliferation during embryogenesis (Tago et al., 2000).

Interestingly, the AMH, a protein broadly reported as an important hormone that acts in the male gonad development in several species, making the reduction of Mullerian ducts in fetal testis, which in females would differentiate into Fallopians tubes and uterus (mammals) (revised Rodriguez-Marí et al., 2005; Manolakou et al., 2006; Ijiri et al., 2008), have divergent edges between both species. For zebrafish AMH interact with FOXO5 and for

human interacts with the ATRX, thus it is possible that AMH works differentially within each species, acting as antagonistic and synergic interactor for zebrafish and humans, respectively.

Moreover, the PPI networks reported here showed other interesting view, that is clearly reported in the edges among proteins that works in nervous system development pathways (SOX3, ATRX and FGF9 for humans), confirming previous reports that nervous system is very important in the sexual development. Moreover, for zebrafish network a lot of proteins have edges with other FGFs proteins (FGF13, FGF16 and FGF8) (data not shown), that in mammals are present in the mammalian X chromosome and usually are important for female development (Wilhelm and Koopman, 2006; O'Meally et al., 2012).

Stopping the SOX maleness pathway and the evolutive role conservation of SOXs

The SOX family is composed of proteins characterized by the HMG box domain, which is involved in DNA binding and are known to work in the developmental stages of the organism (Bowles et al., 2000). The Sry gene, one member of SOX family, is a master-key gene in the male sex determination in therian mammals (Graves and Peichel, 2010). Moreover, it was exhaustively reported an absence of Sry in non-therian animals (Waters et al., 2007). The SOX9a/SOX9 is broadly discussed to be a master sex determiner in non-therian mammals (revised Morrish and Sinclair, 2002; Ijiri et al., 2008). Graves and Peichel (2010) summarized the sex determination in mammals and they reported that the expression of Beta-catenin (beginning of Beta-catenin pathway) is very important to female development. Taken together, we tried to understand how those SOXs here studied work to repress the expression of Beta-catenin or *vice-versa*, getting the shortest paths from SOXs proteins to CTNNBIP1 (Beta-catenin-interacting protein 1), that in mouse interact with Beta-catenin (Tago et al., 2000).

The shortest paths ranging from SOX proteins to CTNNBIP1 at both D-SDD and H-SDD are different according the SOX protein analyzed, but at least two intermediated proteins, the CYP19A1B/CYP19A1 and FOXO5/FOXO3, are reported as conserved across those three paths. As aforementioned, the CYP19 constitute aromatases that convert androgen to estrogen (Trant et al., 2001) and is suggested by gene ontology that FOXO5/FOXO3 could work in the female development. Despite of CTNNBIP1 be reported as suppressor of the linkage of Beta-catenin with WNT, being a important step in the general development and cell proliferation

(Tago et al., 2000), we suggest that probably CYP19A1B/CYP19A1 and FOXO5/FOXO3 could act repressing the activities of SOX proteins binding to the SOXs or directly to the CTNNBIP1. It could allow the female development, a hypothesis to be further experimentally checked.

SOX3 is very important in the brain development and normal brain function in mammals and other vertebrates (revised by Kiefer et al., 2007; Sutton et al., 2011). It is known that SRY evolved from SOX3 in ancestral therian mammals and suffered a neo-functionalization process, becoming an important sex determiner (revised by Mawaribuchi et al., 2012; Graves and Peichel, 2010). Taken together, our D-SDD and H-SDD data are in agreement with the evolution of SRY protein, since SOX3 at D-SDD is also working in the sex determination such as SOX9A. Thus, we suggested that SOX3 in non-therian vertebrates may be very important to sex determination equally comparable to the importance well reported for SOX9A (in fact they have relative similar edges at D-SDD). Probably SOX9A and SOX3 are working together to repress the femaleness or starting the normal male development, at least for zebrafish. In fact, it was observed that SOX3 was expressed ectopically in bipotential gonads in mice transgenic lines, conducting to the XX male sex reversal (Sutton et al., 2011); and is expressed in adult testis and ovary in marsupials (revised by Waters et al., 2007). Anyway, our results support previous reports that probably the SOX3 in therian mammals is more important to the brain development than the male gonads development (a role of other SOXs), and shed light concerning the importance of this protein in sex determination in non-therian vertebrates.

Alternatively, FOXL2, that is broadly known to work in the ovarian determination development in several species, acting as "anti-testis development" (Wang et al., 2004; Hudson et al., 2005; Uhlenhaut and Treier, 2006; Garcia-Ortiz et al., 2009), could interact antagonistically to DMRT1 and/or SOX3, and together with RSPO1, CYP19A1B and FOXO5 may interrupt the development of maleness breaking the SOX pathways at the D-SDD. However, this hypothesis is not applicable to H-SDD, because some of those edges are missing. But for humans, probably the SOX breaking could be intermediated by other factor not observed here.

Unsolved questions

Some interactions here reported are a little bit not clear. For example, it is reported to mammals that the WT1 interacts with SRY (Matsuzawa-Watanabe et al., 2003) and our analysis does not report any interaction between WT1A/WT1 with SOX proteins. Are the actual sub-networks not reporting other proteins that could do these associations? Could the SOXs bind to ATRX before to link to WT1? Similar questions could be formulated for D-SDD linkages? These questions can not be solved analyzing the present sub-networks, but an enrichment with other proteins could clarify such issue.

The aromatases are known to work in the initial stages of sex determination in several species (Morrish and Sinclair, 2002; Ijiri et al., 2008) and in zebrafish CYP19A1A works in the gonad differentiation (Trant et al., 2001; Rodrigues-Marí et al., 2005). However, this protein has just one connection in the D-SDD (data not shown). The point is why the brain aromatase is higher connected than gonad one within this sub-network? Our reciprocal best-hit reported that the brain copy is ortolog of CYP19A in humans, thus, we could suggest that the brain aromatase may be very important to sex determination, such as the gonad one. Anyway, another explanation could be that the relationship between gonad aromatase in the sex determination and differentiation pathways is a regulatory relationship, which is not reported using the present approach.

Conclusion

This reports showed a relative low conservation between zebrafish and human protein-protein networks and this low conservation is also observed when analyzed sub-networks related to vertebrate SDD. Probably it could occur because of evolutionary aspects inherent to the genomic and/or protein evolutionary process may have changed the networks across the evolution, adding or deleting nodes or performing edge rewritings. Concerning the sub-network involved in the sex-determination and differentiation, we conclude that despite to have a relative low conservation, the evolutionary role (sex determination and differentiation) was maintained across the evolution. This is in agreement with the current idea that the sex determination in vertebrates, mainly within fish groups, is very diverge and several factors could influence the fate of a bipotential gonad (Mank et al., 2006; Mank and Avise, 2009).

Moreover, we conclude that in fact the FOXO5/FOXO3 and the aromatases are really important proteins in the vertebrate SDD to be further experimentally studied, because they work like hubs and participate as intermediate proteins at the SOXs-Beta-catenin information flow. In fact, hubs have a great probability of three time more to be essential proteins, with a central role in the network and usually they evolve slowly as well as their connections (revised by Börnke, 2008). Another conclusion about the SDD networks is concerning the role of SOXs proteins, that we suppose to be conserved across the evolution, being the SOX9A and SOX3 equally important to the male sex determination in zebrafish; while the SRY and SOX9 in mammals grabbed that function to therian mammals.

Anyway, this paper reports new perspectives to discuss about the evolutionary aspects of protein-protein interaction networks and bring new candidates to be further investigated by experimental approaches. Moreover, both networks are one of the highest reported until now for both species and will be freely available to the scientific community to be investigated according to different scientific purposes.

References

As referências deste capítulo estão listadas no tópico “7. Referências bibliográficas”.

Figures

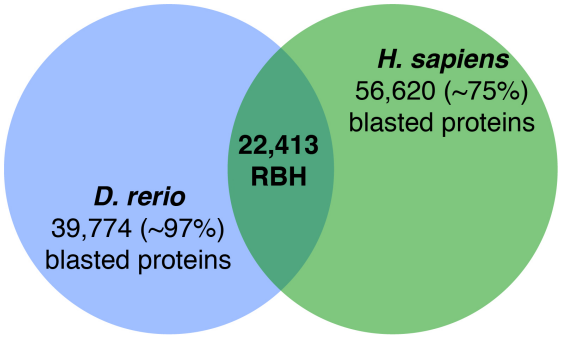


Figure 1 – Venn diagram reporting the results of the reciprocal best-hit between zebrafish and human. RBH, reciprocal best-hit.

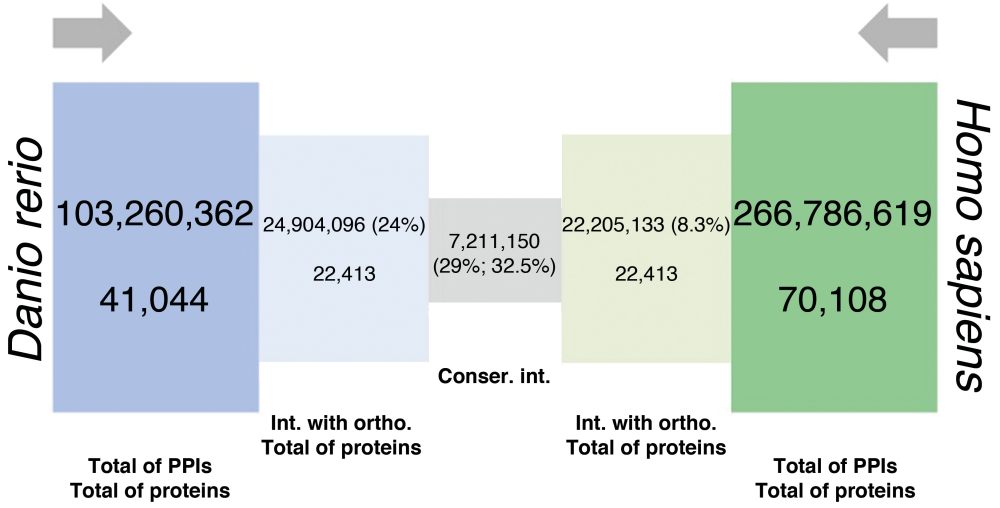


Figure 2 – Description of instances and protein numbers in zebrafish and human PPI networks. Int, interactions; Ortho, orthologs; Conser. int, conserved interactions or interologs. The grey arrows indicate the flow of information. The percentages are relative to the amount of interactions reported in the previous box.

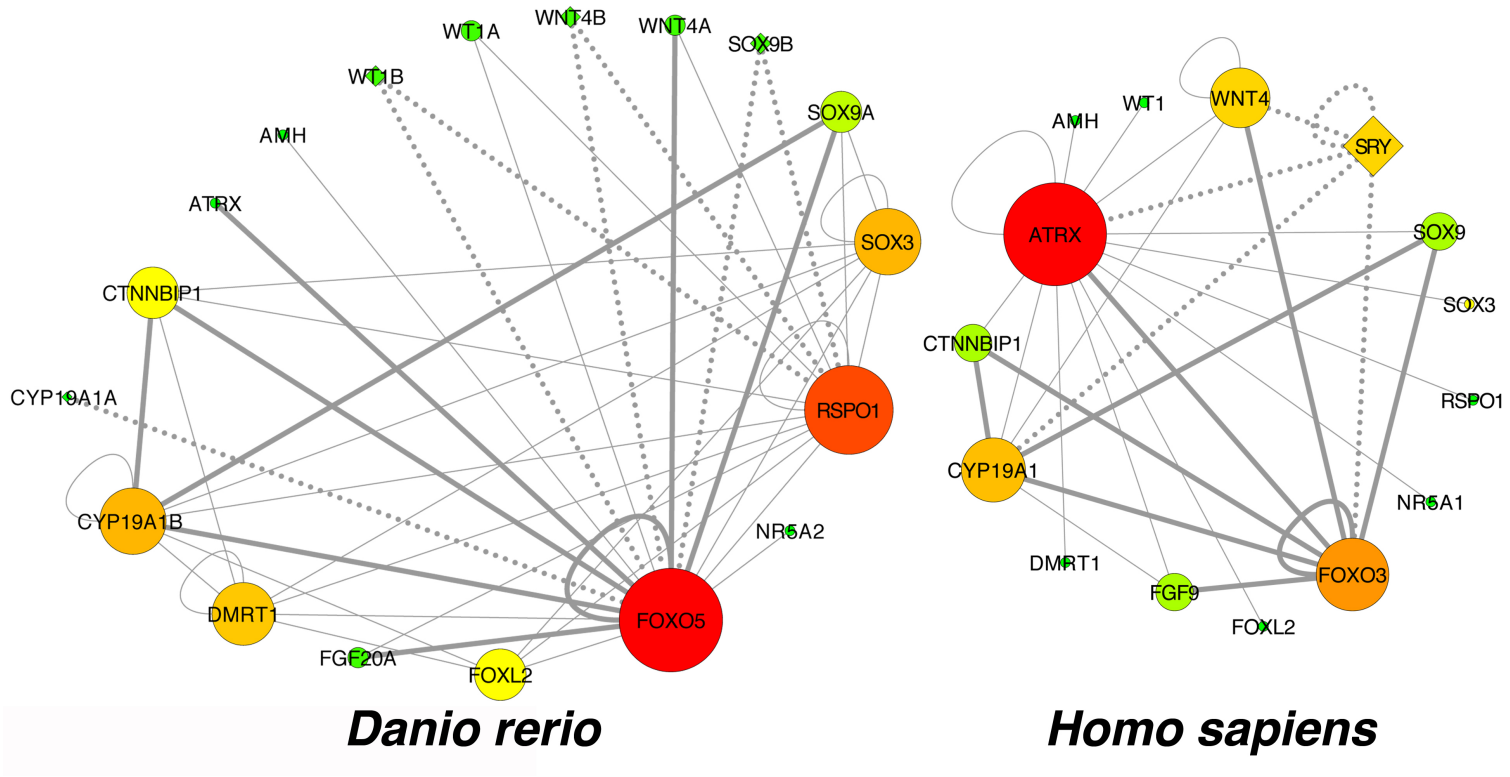


Figure 3 – Graphical representation of D-SDD and H-SDD. The node colors and sizes ranging from red to green and large to small according to decreasing degree of distribution. Circle nodes, proteins with orthologs. Square nodes, proteins without ortholog. Large edges, interologs. Thin edges, non-conserved interactions. Dot edges, non-conserved interactions due to the absence of an ortholog.

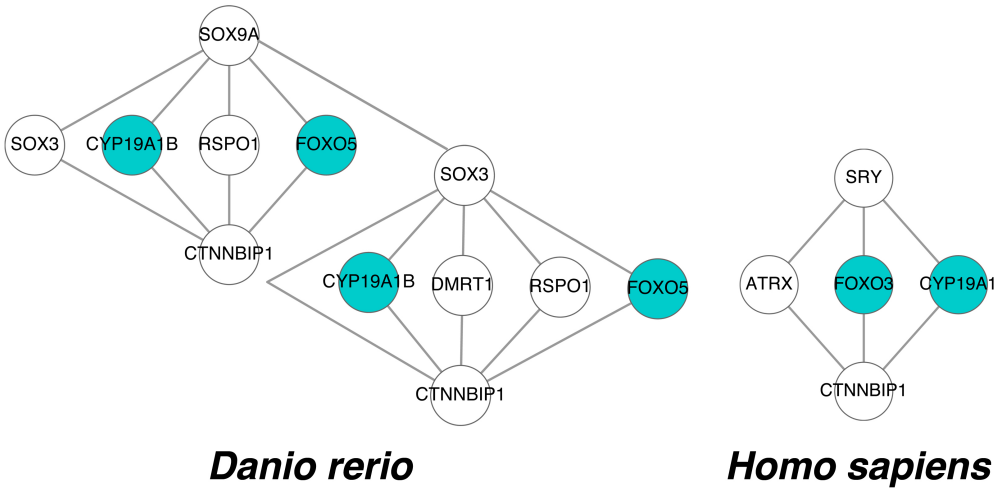


Figure 4 – The shortest paths among SOX proteins (SRY, SOX9a and SOX3) and CTNNBIP1 for D-SDD and H-SDD. Green nodes, conserved intermediate nodes in those paths.

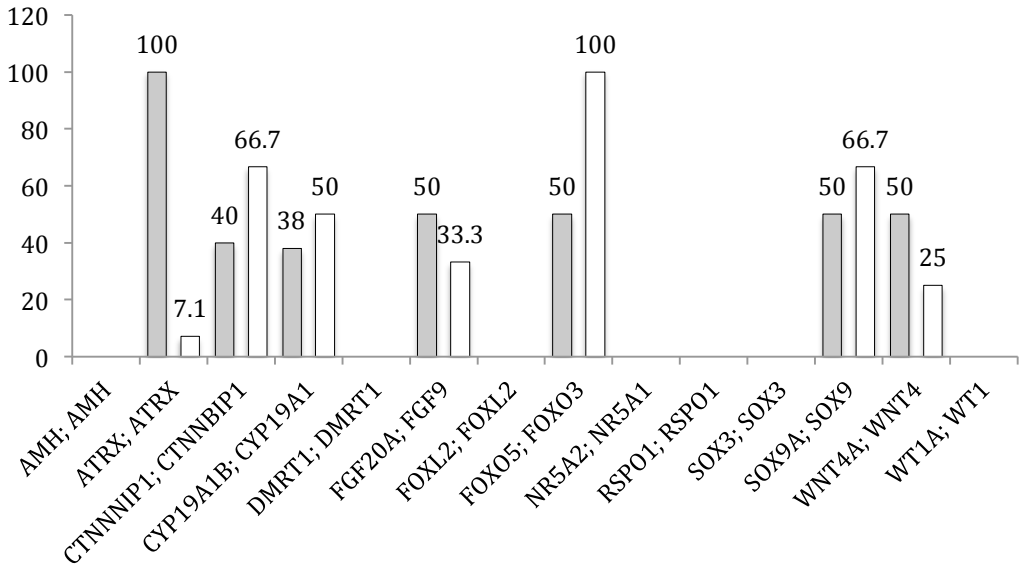


Figure 5 – Graphic showing the percentage of interologs between D-SDD and H-SDD. These values were obtained from the degree of distribution of the proteins with an ortholog and excluding the interactions with non-ortholog proteins. Thus, one value of 25% indicates that 75% of the interactions of that node is not conserved. It is represented the name of proteins for both species, being the first one for zebrafish. Grey bars, percentage for D-SDD. White bars, percentage for H-SDD.

Table 1 – Proteins related to the large sub-network of vertebrate SDD used to build the D-SDD and H-SDD. The proteins between both species are ordered according their orthologies. Their UniProt accession numbers are reported. *, indicate the interactions used to build the D-SDD and H-SDD. -, mean proteins without an ortholog or not included in the SDDs.

Danio rerio		***Homo sapiens***	
Proteins	**UniProt access**	**Proteins**	**UniProt access**
AMER1	F1RDM5	AMER1	Q5JTC6
* AMH	A8E7B8	* AMH	P03971
ATRX	E7F4M2	ATRX	P46100
* ATRX	F1QJ36	* ATRX	H0Y3T0
CTNNB1	F1QGH7	CTNNB1	P35222
* CTNNBIP1	Q9I903	* CTNNBIP1	Q9NSA3
* CYP19A1A	O42145	-	-
* CYP19A1B	Q5U870	* CYP19A1	P11511
DMRT1	A2BFB9	DMRT1	F5GW87
* DMRT1	Q71MM5	* DMRT1	Q9Y5R6
EAF1	Q6IQD2	EAF1	Q96JC9
EAF2	Q6IQU8	EAF2	Q96CJ1
FGF10A	Q8AY90	FGF10	O15520
* FGF20A	Q38PL0	* FGF9	P31371
FGF8A	O57341	FGF8	P55075
* FOXL2	Q1JPV2	* FOXL2	P58012
* FOXO5	Q5TYS1	* FOXO3	O43524
FSHB	Q6TCF7	FSHB	P01225
KLF9	A8E7A9	WT1	H0Y7K5
* NR5A2 (Sf1)	F1QFW6	* NR5A1 (Sf1)	Q13285
* RSPO1	Q6DHR0	* RSPO1	Q2MKA7
* SOX3	Q6EJB7	* SOX3	P41225
* SOX9A	Q9DFH2	* SOX9	P48436
SOX9A	F1Q8I8	SOX9	E7EPD5
* SOX9B	F1QDZ3	-	-
-	-	SRY	Q05066

UNCHARPROT	E7FC99	CTNNB1	B4DGU4
UNCHARPROT	F1QYT6	CTNNB1	C9IZ65
* WNT4A	P47793	* WNT4	B1AJZ6
WNT4A	Q1RLW9	WNT4	P56705
* WNT4B	A3KQ12	-	-
* WT1A	Q9PUT7	* WT1	P19544
* WT1B	Q1L884	-	-
WTAP	Q7SXL7	WTAP	Q15007
WTIP	A8DZE6	WTIP	A6NIX2

Table 2 – The degree of distribution of the proteins of the large sub-network of vertebrate SDD. Degree 1, degree of distribution does not considering the orthologies of interactors. Degree 2, degree of distribution considering the orthologies of interactors.

Danio rerio			***Homo sapiens***		
Proteins	**Degree 1**	**Degree 2**	**Proteins**	**Degree 1**	**Degree 2**
AMER1	1,284	411	AMER1	2,401	466
AMH	520	211	AMH	2,401	466
ATRX	1,072	490	ATRX	1,979	486
ATRX	1,125	523	ATRX	16,891	5,537
CTNNB1	4,429	2,020	CTNNB1	7,012	1,871
CTNNBIP1	5,216	2,333	CTNNBIP1	7,716	2,061
CYP19A1B	10,957	5,559	CYP19A1	2,137	542
DMRT1	10,673	5,302	DMRT1	2,401	466
DMRT1	11,649	5,975	DMRT1	1,214	206
EAF1	20,825	10,889	EAF1	8,349	2,246
EAF2	5,488	2,458	EAF2	1,667	358
FGF10A	11,064	5,571	FGF10	58,058	17,517
FGF20A	1,291	415	FGF9	6,927	1,843
FGF8A	4,574	2,123	FGF8	1,238	211
FOXL2	4,057	1,907	FOXL2	2,401	466
FOXO5	41,000	22,388	FOXO3	6,684	1,800
FSHB	912	281	FSHB	2,401	466
KLF9	1,284	411	WT1	2,401	466
NR5A2	484	177	NR5A1	2,401	466
RSPO1	30,996	16,293	RSPO1	2,059	361
SOX3	17,903	9,530	SOX3	2,401	466
SOX9A	4,114	1,922	SOX9	7,197	1,927
SOX9A	4,511	2,131	SOX9	6,539	1,775
UNCHARPROT	4,310	1,982	CTNNB1	7,012	1,871
UNCHARPROT	4,310	1,982	CTNNB1	2,401	466
WNT4A	946	296	WNT4	13,710	3,954
WNT4A	946	296	WNT4	2,401	466
WT1A	1,623	652	WT1	2,401	466

WTAP	4,768	2,194	WTAP	24,856	8,353
WTIP	2,673	1,250	WTIP	2,401	466

Table 3 – The degree of distribution of the proteins from D-SDD and H-SDD.

Danio rerio		***Homo sapiens***	
Proteins	**Dregree of distribution**	**Proteins**	**Dregree of distribution**
AMH	1	AMH	1
ATRX	1	ATRX	15
CTNNBIP1	5	CTNNBIP1	3
CYP19A1B	8	CYP19A1	7
DMRT1	7	DMRT1	1
FGF20A	2	FGF9	3
FOXL2	5	FOXL2	1
FOXO5	18	FOXO3	8
NR5A2	1	NR5A1	1
RSPO1	14	RSPO1	1
SOX3	8	SOX3	1
SOX9A	4	SOX9	3
WNT4A	2	WNT4	5
WT1A	2	WT1	1

6. Conclusão

A presente tese relatou o desenvolvimento de um método universal para predizer interações entre proteínas e esse método foi aplicado para o estudo de determinação e diferenciação sexual em vertebrados.

Quanto ao método de predição de interações entre proteínas, existe uma grande expectativa de aceitação por parte da comunidade científica, uma vez que este estudo representa uma ferramenta potencial para as mais variadas áreas de pesquisa, podendo abranger aspectos evolutivos, fisiológicos e até clínicos. Como exemplo, pode-se utilizar o método para estudos em biologia fundamental, tal como está relatado no segundo capítulo dessa tese; ou para estudos clínicos e farmacêuticos, onde o método pode ser aplicado para ranquear as proteínas que interagem entre um hospedeiro/patógeno, criando assim um *set* de alvos para desenvolvimento de fármacos. De qualquer forma, UNISPPI pode se tornar uma ferramenta muito útil para estudos em biologia de sistemas.

As análises sobre a rede de interação entre proteínas para a determinação e diferenciação sexual em vertebrados não corrobora completamente as hipóteses prévias que nortearam esse trabalho, no entanto, alguns aspectos das redes são relativamente conservadas. Apesar disso, algumas pesquisas reportam que esses mecanismos são bastante diversos e variáveis, principalmente dentro do grupo teleósteos (Mank et al., 2006; Mank e Avise, 2009). Nesse contexto, os nossos dados estão em relativa concordância com alguns relatos na literatura. Além disso, um novo conjunto de genes candidatos nessas vias foi elencado como sendo importantes alvos a serem investigados por procedimentos experimentais. Contudo, os resultados obtidos abrem perspectivas para novas discussões e trabalhos voltados para o estudo da evolução dos mecanismos de determinação e diferenciação sexual nos vertebrados.

7. Referências bibliográficas

Anashkina A., Kuznetsov E., Esipova N., Tumanyan V. (2007) Comprehensive Statistical Analysis of Residues Interaction Specificity at Protein-Protein Interfaces. Protein. 67: 1060-1077.

Anfinsen C.B. (1973) Principles that govern the folding of protein chains. Science 181: 223-230.

Archakov A.I., Govorun V.M., Dubanov A.V., Ivanov I.D., Veselovsky A.V., Lewi P., Janssen P. (2003) Protein-protein interactions as a target for drugs in proteomics. Proteomics. 3: 380-391.

Arkhipchuk V.V. (1995) Role of chromosomal and genome mutations in the evolution of bony fishes. Hydrobiol J. 31: 55-65.

Aytuna A.S., Gursoy A., Keskin O. (2005) Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. Bioinformatics. 21: 2850-2855.

Bahadur R.P. (2010) A Structural Perspective on Protein–Protein Interactions in Macromolecular Assemblies. In: Protein-Protein Complexes: Analysis, Modeling and Drug Design, Zacharias M (ed). Imperial College Press. p. 25-45.

Barske L.A., Capel B. (2008) Blurring the edges in vertebrate sex determination. Cur Opint Genet Deve. 18: 499-505.

Bellomo N., Carbonaro B. (2011) Toward a mathematical theory of living systems focusing on developmental biology and evolution: a review and perspectives. Phy Life Rev. 8: 1-18.

Ben-Hur A., Noble W.S. (2005) Kernel methods for predicting protein–protein interactions. Bioinformatics. 21: i38-i46.

Bernard P., Harley V.R. (2007) Wnt4 action in gonadal development and sex determination. Int J Biochem Cell Biol. 39: 31-43.

Bock J.R., Gough D.A. (2001) Predicting protein–protein interactions from primary structure. Bioinformatics. 17: 455-460.

Bowles J., Schepers G., Koopman P. (2000) Phylogeny of the SOX Family of Developmental Transcription Factors Based on Sequence and Structural Indicators. Dev Biol. 227: 239-255.

Breiman L. (1996) Bagging predictors. Mach Learn. 24: 123.

Bloom J.D., Drummond D.A., Arnold F.H., Wilke C.O. (2006) Structural determinants of the rate of protein evolution in yeast. Mol Biol Evol, 23: 1751-1761.

Börnke F. (2008) Protein interactions networks. In: Analysis of Biological Networks. Junker B.H., Schreiber F. (eds). John Wiley & Sons. p. 207-232.

Chae L., Lee I., Shin J., Rhee S.Y. (2012) Towards understanding how molecular networks evolve in plants. Cur Opin Plant Biol. 15: 177-1184.

Chen X.W., Jeong J.C. (2009) Sequence-based prediction of protein interaction sites with an integrative method. Bioinformatics, 25: 585-591.

Chen X.W., Liu M. (2005) Prediction of protein-protein interactions using random decision forest framework. Bioinformatics, 21: 4394-4400.

De Bodt S., Proost S., Vandepoele K., Rouzé P., Van de Peer Y. (2009) Predicting protein-protein interactions in Arabidopsis thaliana through integration of orthology, gene ontology and co-expression. BMC Genomics, 10: 288.

Devlin R.H., Nagahama Y. (2002) Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. Aquaculture, 208: 191-364.

Dunker A.K., Brown C.J., Lawson J.D., Iakoucheva L.M., Obradovic Z. (2002) Intrinsic disorder and protein function. Biochemistry, 41: 6573-6582.

Drummond D.A., Bloom J.D., Adami C., Wilke C.O., Arnold F.H. (2005) Why highly expressed proteins evolve slowly. Proc Natl Acad Sci USA, 102: 14338-14343.

Fox A., Taylor D., Slonim D.K. (2009) High throughput interaction data reveals degree conservation of hub proteins. Pac Symp Biocomput, 391-402.

Gandhi T.K.B., Zhong J., Mathivanan S., Karthick L., Chandrika K.N., Mohan S.S., Sharma S., Pinkert S., Nagaraju S., Periaswamy B., et al. (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet, 38: 285-293.

Garcia-Ortiz J.E., Pelosi E., Omari S., Nedorezov T., Piao Y., Karmazin J., Uda M., Cao A., Cole S.W., Forabosco A., Schlessinger D., Ottolenghi C. (2009) Foxl2 functions in sex determination and histogenesis throughout mouse ovary development. BMC Dev Biol, 9: 36.

Graves J.M, Peichel C.L. (2010) Are homologies in vertebrate sex determination due to shared ancestry or to limited options? Gen Biol, 11: 205.

Guo Y., Li M., Pu X., Li G., Guang X., Xiong W. Li J. (2010) PRED_PPI: a server for

predicting protein-protein interactions based on sequence data with probability assignment. BMC Res Notes. 3: 145.

Guo Y., Yu L., Wen Z., Li M. (2008) Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res, 36: 3025-3030.

Han J.D., Dupuy D., Bertin N., Cusick M.E., Vidal M. (2005) Effect of sampling on topology predictions of protein-protein interaction networks. Nat Biotechnol, 23: 839-844.

Hattori R.S., Murai Y., Oura M., Masuda S., Majhi S.K., Sakamoto T., Fernandino J.I., Somoza G.M., Yokota M., Strüssmann C.A. (2012) A Y-linked anti-Müllerian hormone duplication takes over a critical role in sex determination. Proc Natl Acad Sci USA, 109: 2955-2959.

Hudson Q.J., Smith C.A., Sinclair A.H. (2005). Aromatase inhibition reduces expression of FOXL2 in the embryonic chicken ovary. Dev Dyn, 233: 1052-1055.

Hue M., Riffle M., Vert J.P., Noble W.S. (2010) Large-scale prediction of protein-protein interactions from structures. BMC Bioinformatics, 11: 144.

Ijiri S., Kaneko H., Kobayashi T., Wang D-S., Sakai F., Paul-Prasanth B., Nakamura M., Nagahama Y. (2008) Sexual dimorphic expression of genes in gonads during early differentiation of a teleost fish, the Nile tilapia *Oreochromis niloticus*. Biol Reprod, 78: 333-341.

Jaillon O., Aury J-M., Wincker P. (2009) "Changing by doubling", the impact of Whole Genome Duplications in the evolution of eukaryotes. C R Biol, 332: 241-253.

Jansen R., Yu H., Greenbaum D., Kluger Y., Krogan N.J., Chung S., Emili A., Snyder M., Greenblatt J.F., Gerstein M. (2003) A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. Science, 302: 449-453.

Jones S., Thornton J.M. (1996) Principles of protein-protein interactions. Proc Natl Acad Sci USA, 93: 13-20.

Junker B.H. (2008). Networks in Biology. In: Analysis of Biological Networks. Junker B.H., Schreiber F. (eds). John Wiley & Sons. p. 3-12.

Kiefer J.C. (2007) Back to basics: Sox genes. Dev Dyn, 236: 2356-2366.

Kingsford C, Salzberg SL (2008) What are decision trees?. Nat Biotechnol, 26: 1011-1013.

Kumar S., Hedges S.B. (1998) A molecular timescale for vertebrate evolution. Nature,

392917-920.

Larrañaga P., Calvo B., Santana R., Bielza C., Galdiano J., Inza I., Lozano J.A., Armañanzas R., Santafé G., Pérez A., Robles V. (2006) Machine learning in bioinformatics. Brief Bioinfo, 7: 86-112.

Liu X., Liu B., Huang Z., Shi T., Chen Y., Zhang J. (2012) SPPS: a sequence-based method for predicting probability of protein-protein interaction partners. Plos one, 7:e30938.

Li Z.R., Lin H.H., Han L.Y., Jiang L., Chen X., Chen Y.Z. (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res, 34: W32-W37.

Lodé T. (2012) Sex and the origin of genetic exchanges. Trends Evol Biol, 4:1-5.

Lo S.L., Cai C.Z., Chen Y.Z., Chung M.C. (2005) Effect of training datasets on support vector machine prediction of protein-protein interactions. Proteomics, 5: 876-884.

Mank J.E., Avise J.C. (2009) Evolutionary diversity and turn-over of sex determination in teleost fishes. Sex Dev, 3: 60–67.

Mank J.E., Promislow D., Avise J. (2006) Evolution of alternative sex-determining mechanisms in teleost fishes. Biol J Linnean Soc, 87: 83-93.

Manolakou P., Lavranos G., Angelopoulou R. (2006) Molecular patterns of sex determination in the animal kingdom: a comparative study of the biology of reproduction. Reprod Biol Endocrinol, 4: 59.

Martin S., Roe D., Faulon J. (2005) Predicting protein-protein interactions using signature products. Bioinformatics, 21: 218-226.

Matsuda M., Shinomiya A., Kinoshita M., Suzuki A., Kobayashi T., Paul-Prasanth B., Lau E., Hamaguchi S., Sakaizumi M., Nagahama Y. (2007) DMY gene induces male development in genetically female (XX) medaka fish. Proc Natl Acad Sci USA, 104: 3865-3870.

Matsuzawa-Watanabe Y., Inoue J., Semba K. (2003) Transcriptional activity of testis-determining factor SRY is modulated by the Wilms' tumor 1 gene product, WT1. Oncogene, 22: 7900-7904.

Mawaribuchi S., Yoshimoto S., Ohashi S., Takamatsu N., Ito M. (2012) Molecular evolution of vertebrate sex-determining genes. Chromosome Res, 20: 139-151.

Mian I.S., Bradwell A.R., Olson A.J. (1991) Structure, function and properties of antibody binding sites. J Mol Biol, 217: 133-151.

Ming R., Moore P.H. (2007) Genomics of sex chromosomes. Curr Opin Plant Biol, 10: 123-30.

Morrish B.C., Sinclair A.H. (2002) Review Vertebrate sex determination: many means to an end. Reproduction, 124: 447-457.

Mueller T.D. Feigon J. (2002). Solution Structures of UBA Domains Reveal a Conserved Hydrophobic Surface for Protein-Protein Interactions. J Mol Biol, 319: 1243-1255.

Nakamura M. (2010) The mechanism of sex determination in vertebrates-are sex steroids the key-factor? J Exp Zool, 313A: 381-398.

Nanni L., Lumini A. (2006) An ensemble of K-local hyperplanes for predicting protein-protein interactions. Bioinformatics, 22: 1207-1210.

Ofran Y., Rost B. (2003) Analysing Six Types of Protein–Protein Interfaces. J Mol Biol, 325: 377-387.

O'Meally D., Ezaz T., Georges A., Sarre S.D., Graves J.A.M. (2012) Are some chromosomes particularly good at sex? Insights from amniotes. Chrom Res, 20: 7-19.

Otto S.P. (2009) The evolutionary enigma of sex. Am Nat, 174: S1-S14.

Padlan E.A. (1990) On the nature of antibody combining sites: Unusual structural features that may confer on these sites an enhanced capacity for binding ligands Proteins, 7: 112-124.

Park Y., Marcotte E.M. (2011) Revisiting the negative example sampling problem for predicting protein-protein interactions. Bioinformatics, 27: 3024-3028.

Pavlopoulos G.A., Secrier M., Moschopoulos C.N., Soldatos T.G., Kossida S., Aerts J., Schneider R., Bagos P.G. (2011) Using graph theory to analyze biological networks. Biodata Mining, 4: 10.

Pitre S., Alamgir M.D., Green J.R., Dumontier M., Dehne F., Golshani A. (2008) Computational Methods For Predicting Protein-Protein Interactions. Adv Biochem Eng Biotechnol, 110: 247-267.

Pough F.H., Janis C.M., Heiser J.B. (2008) A Vida dos Vertebrados. 4 ed. Ateneu. 684 p.

Quinlan J.R. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann. 302 p.

Ren X., Wang Y.C., Wang Y., Zhang X.S., Deng N.Y. (2011) Improving accuracy of protein-protein interaction prediction by considering the converse problem for sequence representation. BMC Bioinformatics, 12: 409.

Res I., Mihalek I., Lichtarge O. (2005) An evolution based classifier for prediction of protein

interfaces without using protein structures. Bioinformatics, 21: 2496-2501.

Rhodes D.R., Tomlins S.A., Varambally S., Mahavisno V., Barrette T., Kalyana-Sundaram S., Ghosh D., Pandey A., Chinnaiyan A.M. (2005) Probabilistic model of the human protein-protein interaction network. Nature Biotechnol, 23: 951-959.

Rodríguez-Marí A., Yan Y-L., Bremiller R.A., Wilson C., Cañestro C., Postlethwait J.H. (2005) Characterization and expression pattern of zebrafish Anti-Müllerian hormone (Amh) relative to sox9a, sox9b, and cyp19a1a, during gonad development. Gene Expre Patt, 5: 655-667.

Roy S., Martinez D., Platero H., Lane T., Werner-Washburne M. (2009) Exploiting amino acid composition for predicting protein-protein interactions. Plos one, 4:e7813.

Sandra G-E., Norma M-M. (2010) Sexual determination and differentiation in teleost fish. Rev Fish Biol Fisheries, 20: 101-121.

Shen J., Zhang J., Luo X., Zhu W., Yu K., Chen K., Li Y., Jiang H. (2007) Predicting protein–protein interactions based only on sequences information. Proc Natl Acad Sci USA, 104: 4337-4341.

Shi M.G., Xia J.F., Li X.L., Huang D.S. (2010) Predicting protein-protein interactions from sequence using correlation coefficient and high-quality interaction dataset. Amino Acids, 38: 891-899.

Shoemaker B.A., Panchenko A.R. (2007a) Deciphering protein–protein interactions. Part I. Experimental techniques and databases. Plos Comput Biol, 3:e42.

Shoemaker B.A., Panchenko A.R. (2007b) Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. Plos Comput Biol, 3:e43.

Simone F., Polak P.E., Kaberlein J.J., Luo R.T., Levitan D.A., Thirman M.K. (2001) EAF1, a novel ELL-associated factor that is delocalized by expression of the MLL-ELL fusion protein. Blood, 98:201-209.

Skrabanek L., Saini H.K., Bader G.D., Enright A.J. (2008) Computational Prediction of Protein-Protein Interactions. Mol Biotechnol, 38: 1-17.

Smialowski P., Pagel P., Wong P., Brauner B., Dunger I., Fobo G., Frishman G., Montrone C., Rattei T., Frishman D., Ruepp A. (2010) The Negatome Database: a reference set of non-interacting protein pairs. Nucleic Acids Res, 38: D540-544.

Smith G.R., Sternberg M.J. (2002) Prediction of protein-protein interactions by docking methods. Curr Opin Struct Biol, 12: 28-35.

Smoot M., Ono K., Ruscheinski J., Wang P-L., Ideker T. (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics, 27: 431-432.

Stark C., Breitkreutz B.J., Chatr-Aryamontri A., Boucher L., Oughtred R., Livstone M.S., Nixon J., Van Auken K., Wang X., Shi X., Reguly T., Rust J.M., Winter A., Dolinski K., Tyers M. (2011) The Biogrid Interaction Database: 2011 update. Nucleic Acids Res, 39: D698-D704.

Steuer R., López G.Z. (2008) Global Network Properties. In: Analysis of Biological Networks. Junker B.H., Schreiber F. (eds). John Wiley & Sons. p. 31-63.

Sutton E., Hughes J., White S., Sekido R., Tan J., Arboleda V., Rogers N., Knower K., Rowley L., Eyre H., et al. (2011) Identification of SOX3 as an XX male sex reversal gene in mice and humans. J Clin Invest, 121: 328-341.

Swaisgood H.E. (2005) The importance of disulfide bridging. Biotechnol Adv, 23: 71-73.

Tago K., Nakamura T., Nishita M., Hyodo J., Nagai S., Murata Y., Adachi S., Ohwada S., Morishita Y., Shibuya H., Akiyama T. (2000) Inhibition of Wnt signaling by ICAT, a novel beta-catenin-interacting protein. Genes Dev, 14: 1741-1749.

Taylor J.S., Raes J. (2004) Duplication and divergence: The evolution of new genes and old ideas. Annual Rev Gen, 38: 615-643.

Trant JM., Gavasso S., Ackers J., Chung B.C., Place AR (2001) Developmental expression of cytochrome P450 aromatase genes (CYP19a and CYP19b) in zebrafish fry (*Danio rerio*). J Exp Zool, 290: 475-483.

Uhlenhaut N.H., Treier M. (2006) Foxl2 function in ovarian development. Mol Genet Metab, 88: 225-234.

UniProt Consortium (2011) Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res, 39: D214-D219.

Valente G.T., Acencio M.L., Martins C., Lemke N. (*in preparation*) The development of a universal *in silico* predictor of protein-protein interactions. Submitted to Molecular System Biology.

Venkatesh B. (2003) Evolution and diversity of fish genomes. Curr Opin Genet Dev, 13: 588-592.

Visa S., Ralescu A. (2005) Issues in mining imbalanced data sets - a review paper. Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference, 67-73.

Waters P.D., Wallis M.C., Graves J.A.M. (2007) Mammalian sex--Origin and evolution of the Y chromosome and SRY. Semin Cell Dev Biol, 18: 389-400.

Wang D., Kobayashi T., Zhou L., Nagahama Y. (2004) Molecular cloning and gene expression of Foxl2 in the Nile tilapia, *Oreochromis niloticus*. Biochem Biophys Res Commum, 320: 83-89.

Wilcoxon F. (1947) Probability tables for individual comparisons by ranking methods. Biometrics, 3: 119-122.

Wilhelm D., Koopman P. (2006) The makings of maleness: towards an integrated view of male sexual development. Nature Rev Genet, 7: 620-631.

Witten I.H., Frank E., Hall M.A. (2011) Data Mining: Practical Machine Learning Tools and Techniques. 3 ed. Morgan Kaufmann. 629 p.

Xiao W., Zhang Q., Jiang F., Pins M., Kozlowski J.M., Wang Z. (2003) Suppression of prostate tumor growth by U19, a novel testosterone-regulated apoptosis inducer. Cancer Res, 63: 4698-4704.

Yoshimoto S., Okada E., Umemoto H., Tamura K., Uno Y., Nishida-Umehara C., Matsuda Y., Takamatsu N., Shiba T., Ito M. (2008) A W-linked DM-domain gene, *DM-W*, participates in primary ovary development in *Xenopus laevis*. Proc Natl Acad Sci USA, 105: 2469-2474.

Yu C.Y., Chou L.C., Chang D.T. (2010a) Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. BMC Bioinformatics, 11: 167.

Yu J., Guo M., Needham C.J., Huang Y., Cai L., Westhead D.R. (2010b) Simple sequence-based kernels do not predict protein-protein interactions. Bioinformatics, 26: 2610-2614.

Zaki N., Lazarova-Molnar S., El-Hajj W., Campbell P. (2009) Protein-protein interaction based on pairwise similarity. BMC Bioinformatics, 10: 150.

Zhang L.V., Wong S.L., King O.D., Roth F.P. (2004) Predicting co-complexed protein pairs using genomic and proteomic data integration. BMC Bioinformatics, 5:38.

Zhao D., McBride D., Nandi S., McQueen H.A., McGrew M.J., Hocking P.M., Lewis P.D., Sang H.M., Clinton M. (2010) Somatic sex identity is cell autonomous in the chicken. Nature, 464: 237-242.

Zhou Y., Zhou Y.S., He F., Song J., Zhang Z. (2012) Can simple codon pair usage predict protein-protein interaction?. Mol Biosyst, 8: 1396-1404.

Zinman G.E., Zhong S., Bar-Joseph Z. (2012) Biological interaction networks are conserved at the module level. BMC Systems Biology, 5:134.

8. Anexos

8.1 Anexo 1 – Trabalhos paralelos não relacionados ao tema da tese publicados durante o período de doutorado

WASKO, A.P. , MARTINS, C. , PINHAL, D., **VALENTE, G.T.**, SILVA, J.R, MOTA, L.S., GOMES, M.L., CAMPOS, M.L. Experimentando genética: Atividades de cultura científica e de inclusão social para o ensino médio público. Aceito na revista Genética na Escola (http://geneticanaescola.com.br/).

VALENTE, G. T., VITORINO, C., CABRAL-DE-MELLO, D., OLIVEIRA, C, SOUSA I, MARTINS, C, VENERE, P. Comparative cytogenetics of ten species of cichlid fishes (Teleostei, Cichlidae) from the Araguaia River system, Brazil, by conventional cytogenetic methods. Comparative Cytogenetics, v.6, p.163 - 181, 2012.

MARTINS, C., CABRAL DE MELO, D. C., **VALENTE, G. T.**, MAZZUCHELLI, J., Oliveria S. G. Cytogenetic Mapping and its Contribution to the Knowledge of Animal Genomes In: Advances in Genetics Research ed. Hauppauge: Nova Science Publisher, 2011, v.4, p. 1-82.

VALENTE, G. T., MAZZUCHELLI, J., FERREIRA, I. A., POLETTO, A. B., FANTINATTI, B.E.A, MARTINS, C. Cytogenetic Mapping of the Retroelements Rex1, Rex3 and Rex6 among Cichlid Fish: New Insights on the Chromosomal Distribution of Transposable Elements. Cytogenetic and Genome Research, v.133, p.34 - 42, 2011.

FANTINATTI, B. E. A., MAZZUCHELLI, J., **VALENTE, G. T.**, CABRAL-DE-MELLO, D. C., MARTINS, C. Genomic content and new insights on the origin of the B chromosome of the cichlid fish Astatotilapia latifasciata. Genetica, v.139, p.1273 - 1282, 2011.

VITORINO, C. A., SOUZA, I. L., ROSA, J. N., **VALENTE, G. T.**, MARTINS, C., VENERE, P. C. Molecular cytogenetics and its contribution to the understanding of the chromosomal diversification in Hoplias malabaricus (Characiformes). Journal of Fish Biology. , v.78, p.1239 - 1248, 2011.

GROSS, M. C., SCHNEIDER, C. H., **VALENTE, G. T.**, MARTINS, C., Feldberg, E. Variability of 18S rDNA locus among fishes: chromosomal rearrangements. Journal of Fish Biology, v.76, p.1117 - 1127, 2010.

GROSS, M. C., SCHNEIDER, C. H., **VALENTE, G. T.**, PORTO, J. I. R., MARTINS, C., FELDBERG, E. Comparative Cytogenetic Analysis of the Genus Symphysodon (Discus Fishes, Cichlidae): Chromosomal Characteristics of Retrotransposons and Minor Ribosomal DNA. Cytogenetic and Genome Research, v.127, p.43 - 53, 2009.

VALENTE, G. T., SCHNEIDER, C. H., GROSS, M. C., FELDBERG, E., MARTINS, C. Comparative cytogenetics of cichlid fishes through genomic in-situ hybridization (GISH) with emphasis on Oreochromis niloticus. Chromosome Research, v.17, p.791 - 799, 2009.

TEIXEIRA, W. G., FERREIRA, I. A., CABRAL-DE-MELLO, D. C., MAZZUCHELLI, J., **VALENTE, G. T.**, PINHAL, D., POLETTO, A. B., VENERE, P. C., MARTINS, C. Organization of Repeated DNA Elements in the Genome of the Cichlid Fish Cichla kelberi and Its Contributions to the Knowledge of Fish Genomes. Cytogenetic and Genome Research, v.125, p.224 - 234, 2009.

8.2 Anexo 2 – Trabalhos paralelos relacionados ao tema da tese, ou envolvidos com análises de bioinformática, produzidos durante o período de doutorado

GERALDO, M. T., **VALENTE, G. T.**, BRAZ, A. S. K., MARTINS, C. The discovery of FOXL2 paralogs in coelacanth and tetrapod genomes reveals an ancient duplication in vertebrates. Aceito na revista Heredity (http://www.nature.com/hdy/index.html).

NAKAJIMA, R. T., CABRAL-DE-MELLO, D. C., **VALENTE, G. T.**, VENERE, P. C., MARTINS, C. Evolutionary dynamics of rRNA gene clusters in cichlid fish. BMC Evolutionary Biology, v.12, p.198, 2012.

CABRAL-DE-MELLO, D. C., **VALENTE, G. T.**, NAKAJIMA, R. T., MARTINS, C. Genomic organization and comparative chromosome mapping of the U1 snRNA gene in cichlid fish, with an emphasis in Oreochromis niloticus. Chromosome Research, v.20, p.279 - 292, 2012.

MARTINS, C., CABRAL DE MELO, D. C., **VALENTE, G. T.**, MAZZUCHELLI, J., Oliveria S. G., PINHAL, D. Animal Genomes Under the Focus of Cytogenetics. New York: Nova Science Publisher, 2011, v.1. p.160.

9. Apêndices e material suplementar

Todo o material suplementar está incluído no CD que acompanha a tese.