João de Abreu e Tôrres

Protein Classification Tool: Uma ferramenta para anotação de proteínas utilizando bases secundárias

Dissertação apresentada ao Departamento de Ciência da Computação da UFMG como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Orientador:
Sérgio Vale Aguiar Campos

Co-orientador:
José Miguel Ortega

Universidade Federal de Minas Gerais
Curso de Pós-Graduação em Ciência da Computação

Agosto 2006 - Belo Horizonte

Agradecimentos

Em primeiro lugar, agradeço aos professores Sérgio Campos e José Miguel Ortega pela orientação ao longo desse trabalho, sem a qual ele não seria possível.

Aos colegas de laboratório que contribuíram com o desenvolvimento do trabalho, Maurício Mudado e Estevam Bravo pela ajuda com o tratamento das bases de dados, Adriano Barbosa pela montagem das bases e implementação preliminar da funcionalidade de comparação de domínios, Daniela Barbosa pela orientação a respeito das técnicas de análise filogenética e Alessandra Campos pela orientação em diversas etapas do trabalho.

A Adriano Gomes, João Andrade Neto, Mariana Faria e Maurício Mudado pela revisão do texto.

E, enfim, aos amigos e a minha família, que me apoiaram ao longo de todo o tempo.

Um sincero muito obrigado a todos vocês!

Resumo

Apesar do grande desenvolvimento das técnicas de seqüenciamento genético (digitalização do material genético) observado nos últimos tempos e do conseqüente crescimento exponencial dos bancos de dados de seqüências, a capacidade de análise dessas seqüências não conseguiu acompanhar esse desenvolvimento. Visto que seqüências não identificadas ou com identificação incorreta são de limitada utilidade científica, este trabalho tem como objetivo tratar esse problema por meio da criação de uma ferramenta a “Protein Classification Tool” (PCT). A PCT implementa as seguintes funcionalidades: comparação usando BLAST com diversas bases secundárias (GOA, COG/KOG e bases do CGAP) e o NR, análise de domínios e análise filogenética. Neste trabalho ainda se define um “framework” para permitir a extensibilidade da ferramenta por meio da adição de bases secundárias e novas funcionalidades, de modo a permitir que a ela se mantenha atualizada, acompanhando os avanços no processo de anotação. Mediante a integração de diversas funcionalidades de anotação, a PCT facilita o processo de anotação (identificação) de seqüências genéticas ao mesmo tempo que permite a obtenção de melhores resultados.

Abstract

Despite the great advances in genetic sequencing technology – digitization of genetic material – observed recently, which resulted in an exponential growth of the amount of data in sequence databases, the ability to properly analyze these sequences could not match this growth. As sequences that lack functional annotation (identification) or with incorrect annotation are of limited use to researchers, this work targets this problem by creating a tool, “Protein Classification Tool” (PCT). The PCT implements the following functionalities: BLAST sequence comparison against several secondary databases (GOA, COG/KOG and CGAP databases), domain analysis and phylogenetic analysis. Moreover, this work defines a framework so that other secondary databases and annotation functionalities may be added to the tool, in order to keep it up-to-date with new annotation techniques. The integration of several annotation functionalities in PCT eases the process of sequence annotation at the same time that allows the obtention of more accurate results

Sumário

Índice de Figuras

Índice de Tabelas

Lista de Abreviaturas

BLAST – Basic Local Alignment Tool

CGAP – Cancer Genome Anatomy Project

COG – Cluster of Orthologous Groups

KEGG – Kyoto Encyclopedia of Genes and Genomes

KOG – Cluster of Orthologous Groups

NCBI – National Center for Biotechnology Information

PCT – Protein Classification Tool

1. Introdução

Nos últimos anos, tem-se observado um grande desenvolvimento nas técnicas de seqüenciamento genético que refletiu no crescimento exponencial do número de seqüências depositadas em bancos de seqüências públicos[1] (Figura 1) e nas centenas de projetos de seqüenciamento de genomas que vêm surgindo. Atualmente, mais de 1600 projetos de seqüenciamento de genomas estão em andamento, e genomas de mais de 400 organismos, entre eucariotos e procariotos, já foram completamente seqüenciados, inclusive o genoma humano[2] (Ainscough et al., 1998; Celniker, 2000; Collins et al., 2001; Venter et al., 2001; Yakunin et al., 2004).

O grande número de seqüências e os progressos em genômica trazem a promessa de contribuir significativamente para melhorar o conhecimento sobre a biologia e a evolução dos organismos, fornecendo novas pistas sobre conteúdo de genes, sua regulação e funcionalidade e os processos evolutivos que originaram os diferentes genomas (Pollock, 2002; Miller et al., 2004). Além disso, há uma expectativa de que o término do seqüenciamento dos vários genomas seja acompanhado da descoberta de alvos para drogas de combate a várias doenças (Nemoto, 1998; Kramer e Cohen, 2004). Entretanto, a realização de tais promessas tem sido, até o momento, limitada pela deficiência na interpretação dos dados fornecidos pelo seqüenciamento. Ou seja, os métodos de análise disponíveis não têm conseguido acompanhar a velocidade em que os dados são gerados pelo seqüenciamento, e o resultado é uma imensa quantidade de dados se convertendo em pouca ou nenhuma informação biológica relevante.

Seqüências de DNA são a matéria-prima do trabalho em genômica, mas o objetivo principal e definitivo de todos os esforços empregados em seqüenciamento é descobrir funções moleculares (bioquímicas) e celulares de todos os produtos gênicos codificados por essas seqüências. A interpretação da informação contida nas seqüências genéticas, chamada de anotação, é uma tarefa não trivial composta de

[1] GenBank - http://www.ncbi.nlm.nih.gov/Genbank/
[2] Genomes OnLine Database - http://www.genomesonline.org/

várias etapas. O processo da anotação tem sido objeto de diversas pesquisas e neste trabalho propomos o desenvolvimento de uma ferramenta para tratar desse problema.

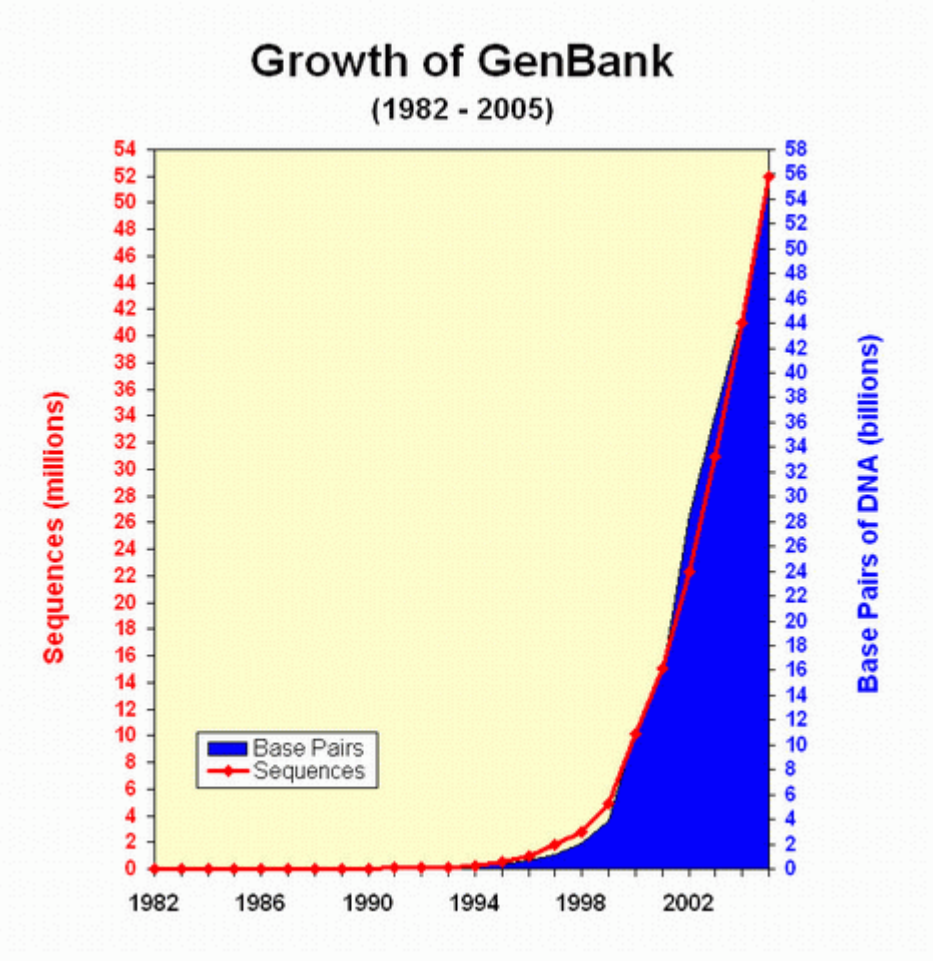


Figura 1 – Gráfico representando o crescimento da base de seqüências GenBank[3]

1.1. Motivação

Pesquisas recentes têm mostrado que para se prover anotação de qualidade é necessária a combinação de diversos métodos, ao contrário de anos atrás em que o

[3] Figura retirada de http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html, versão de 7 de março de 2006

único recurso disponível eram comparações de similaridade em bancos de seqüências não classificados (Mathé et al, 2002).

Enquanto vários métodos de anotação vão se tornando mais populares é preciso ter ferramentas que disponibilizem implementações desses métodos, de modo a permitir seu efetivo uso e de maneira prática na anotação de seqüências.

Visto isso, tem-se a necessidade de uma ferramenta que permita realizar uma anotação de maneira rápida e confiável. Agregando as diferentes funcionalidades atualmente existentes para anotação, e dando ao usuário meios diversos de aperfeiçoar o resultado. Mesmo que esse usuário não tenha profundo conhecimento sobre os diversos meios de anotação e como utilizá-los corretamente.

1.2. Objetivos

Este trabalho descreve o desenvolvimento de uma ferramenta de anotação de seqüências com os seguintes objetivos:

- Reunir numa única ferramenta as principais funcionalidades atualmente em uso no processo de anotação, melhorando tanto a velocidade quanto a qualidade da anotação.

- Criar uma estrutura na qual seja possível agregar facilmente novas funcionalidades, à medida que o processo de anotação evolui, de modo a manter a ferramenta atualizada e funcional.

1.3. Contribuições

A ferramenta desenvolvida neste trabalho gerou as seguintes contribuições:

- Ambiente para execução do BLAST com a base primária NR e diversas bases secundárias e devido tratamento de seu resultado, provendo facilidade de identificação e classificação funcional de seqüências

- Obtenção e tratamento das seguintes bases de dados secundárias:

    - CGAP Biocarta
    - CGAP Kegg
    - COG
    - KOG
    - GOA

- Funcionalidade de análise de estrutura de domínios

- Funcionalidade de geração de árvore de filogenética

- Estrutura que permite a adição de novas funcionalidades e bases secundárias de maneira integrada à ferramenta

2. O fluxo de informação nas células e as seqüências biológicas

2.1. Fluxo de informação nas células

A informação hereditária completa, sobre a total constituição de um organismo é armazenada numa molécula chamada de “DNA” (“deoxyribonucleic acid”, ou ácido desoxirribonucléico, ADN, em português), presente em todas as células de um organismo. Toda vez que uma célula se reproduz, todo o seu DNA é duplicado em um processo chamado de “replicação do DNA” (Figura 2). Já o conjunto de todo o DNA de um organismo é chamado de “genoma”. Seqüenciar (digitalizar) e devidamente interpretar os genomas dos organismos mais relevantes é um dos grandes desafios da ciência atual. Medicina, agricultura e diversas indústrias dependem do conhecimento genômico para desenvolver medicamentos, modificar características específicas em plantas e animais e compreender o relacionamento entre espécies.

O conhecimento que se tem sobre a informação contida no DNA é ainda restrito. Mesmo que nos últimos 50 anos tenha-se avançado consideravelmente na compreensão dessa informação, há muito sobre o qual ainda não se sabe, por exemplo, sobre a função da grande parte do DNA de eucariotos que não codifica proteínas. Por outro lado, os dados propriamente ditos são de composição simples, consistindo de apenas 4 tipos de nucleotídeos, que são referenciados pelas bases nitrogenadas que os diferenciam: adenina, citosina, guanina e timina, ou, respectivamente, A, C, G e T. O DNA é geralmente encontrado como uma molécula na forma de dupla hélice, mas em geral, quando se trata dos dados, trata-se apenas uma das fitas por vez. Uma seqüência de DNA é representada por uma seqüência de texto contendo as letras que representam as bases dessa fita. Além das 4 letras que codificam os diferentes tipos de bases, são usadas mais 11 letras para especificar ambigüidade entre as bases, como visto na Tabela 1.

Símbolo	Nucleotídeo
R	A ou G
Y	C ou T
W	A ou T
S	G ou C
K	G ou T
M	A ou C
B	C, G, ou T
D	A, G, ou T
H	A, C, ou T
V	A, C, ou G
N	A, C, G, ou T

Tabela 1 – Ambigüidade de nucleotídeos

Na estrutura da molécula de DNA, cada base em uma das fitas corresponde a uma outra base na outra fita, sendo que as bases são ligadas aos pares, especificamente, de A-T e C-G. A estrutura de dupla fita do DNA permite não só que a molécula seja mais estável, mas também funciona como um dispositivo de correção de erro no caso de dano a alguma base, causado, por exemplo, pelo excesso de radiação ultravioleta do Sol.

Ao longo do DNA estão codificados os genes. Eles são unidades hereditárias dos organismos, presentes no DNA. Apesar de poderem ser definidos de diversas maneiras, nesse trabalho tratamos os genes como unidades do DNA que contêm instruções para a codificação de uma proteína, apesar de alguns deles produzirem RNAs que não codificam proteínas.

A principal função do DNA é armazenar toda informação genética de um organismo. Toda vez que uma célula se divide, seu DNA é duplicado num processo chamado de “replicação” e a produção de proteínas na célula se dá a partir da cópia de uma seqüência do DNA para o RNA, em um processo chamado de “transcrição” (Figura 2).

A molécula do RNA é bastante semelhante à do DNA, entretanto, no RNA encontramos a base “uracila” (U) no lugar da “timina” (T). Entre outras diferenças, o RNA é, em geral, encontrado como uma molécula de fita simples.

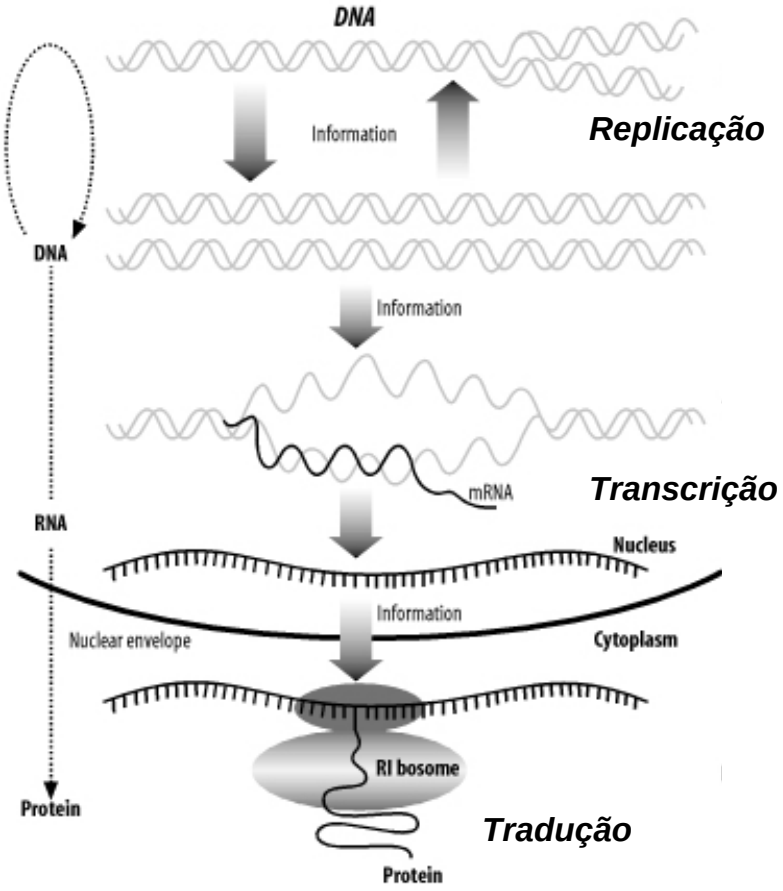


Figura 2 – Figura representando o fluxo de informação na célula[4]

O RNA pode ser de diversos tipos e apresentar diversas funções, entretanto, RNAs que correspondem a genes que codificam proteínas são os RNA mensageiros, ou mRNA.

Entre outras funções, as proteínas são constituintes estruturais do “maquinário” da célula. Elas são moléculas quimicamente diferentes do DNA e RNA, pois são compostas de aminoácidos ao invés de ácidos nucléicos. Proteínas têm a propriedade

[4] Figura retirada de Bedell et al, 2003

de se “dobrar” em formas tridimensionais bastante específicas, que dependem de sua seqüência de aminoácidos. Assim, a seqüência de aminoácidos determina a forma de uma proteína e a forma determina sua função, de modo que há proteínas que desempenham as mais diversas funções num organismo. Com isso, vemos que enquanto, nesse contexto, o DNA e o RNA são utilizados principalmente para armazenamento e transporte de informações, as proteínas são o resultado desse processo, mostrando-se responsáveis por inúmeros processos no organismo.

Os aminoácidos que compõem as proteínas, por sua vez, são codificados, cada um, por 3 nucleotídeos. Uma vez que são 4 os tipos de nucleotídeos, existem 64 combinações possíveis de códons (grupos de 3 nucleotídeos). Entretanto, as 64 combinações, conforme o código genético (Figura 3), codificam apenas 20 aminoácidos (Tabela 2), sendo que um deles, a Metionina indica o início da codificação de uma proteína (“start”) e outros 3 códons indicam o final dessa codificação (“stop”). Assim, vários desses aminoácidos são codificados redundantemente por mais de um tipo de códon. A Tabela 2 lista os 20 aminoácidos e suas respectivas representações, e a Figura 3 ilustra a conversão de nucleotídeos em aminoácidos usando o código genético padrão.

	T	C	A	G
T	TTT Phe (F) TTC " TTA Leu (L) TTG "	TCT Ser (S) TCC " TCA " TCG "	TAT Tyr (Y) TAC TAA Ter TAG Ter	TGT Cys (C) TGC TGA Ter TGG Trp (W)
C	CTT Leu (L) CTC " CTA CTG "	CCT Pro (P) CCC " CCA " CCG "	CAT His (H) CAC " CAA Gin (Q) CAG "	CGT Arg (R) CGC " CGA " CGG "
A	ATT Ile (I) ATC " ATA " ATG Met (M)	ACT Thr (T) ACC " ACA " ACG "	AAT Asn (N) AAC " AAA Lys (K) AAG "	AGT Ser (S) AGC " AGA Arg (R) AGG "
G	GTT Val (V) GTC " GTA " GTG "	GCT Ala (A) GCC " GCA " GCG "	GAT Asp (D) GAC " GAA Glu (E) GAG "	GGT Gly (G) GGC " GGA " GGG "

Figura 3 – Código genético padrão[5]

[5] Figura retirada de Bedell et al, 2003

Nome	Representação
Glicina ou Glicocola	Gly, Gli ou G
Alanina	Ala ou A
Leucina	Leu ou L
Valina	Val ou V
Isoleucina	Ile ou I
Prolina	Pro ou P
Fenilalanina	Phe ou Fen
Serina	Ser ou S
Treonina	Thr, The ou T
Cisteina	Cys, Cis ou C
Tirosina	Tyr, Tir ou Y
Asparagina	Asn ou N
Glutamina	Gln ou Q
Aspartato ou Ácido aspártico	Asp ou D
Glutamato ou Ácido glutâmico	Glu ou E
Arginina	Arg ou R
Lisina	Lys, Lis ou K
Histidina	His ou H
Triptofano	Trp, Tri ou W
Metionina	Met ou M

Tabela 2 – Principais aminoácidos

As proteínas ainda apresentam estruturas internas chamadas “domínios”. Domínios de proteínas podem ser vistos como um trecho da proteína com uma função ou estrutura distinta (freqüentemente ambas). No geral, quando se vê uma unidade estrutural específica dentro de uma proteína, ela costuma ter uma função específica associada a ela. Assim, os domínios determinam características específicas de cada proteína e uma proteína pode conter um ou mais domínios (Figura 4).

2.2. Mutação, evolução e homologia

O pressuposto para sistemas de identificação e anotação de seqüências baseados em similaridade é que os organismos estão evolutivamente relacionados, e derivam, por meio de sucessivas mutações, de um ancestral comum (Bedell et al, 2003).

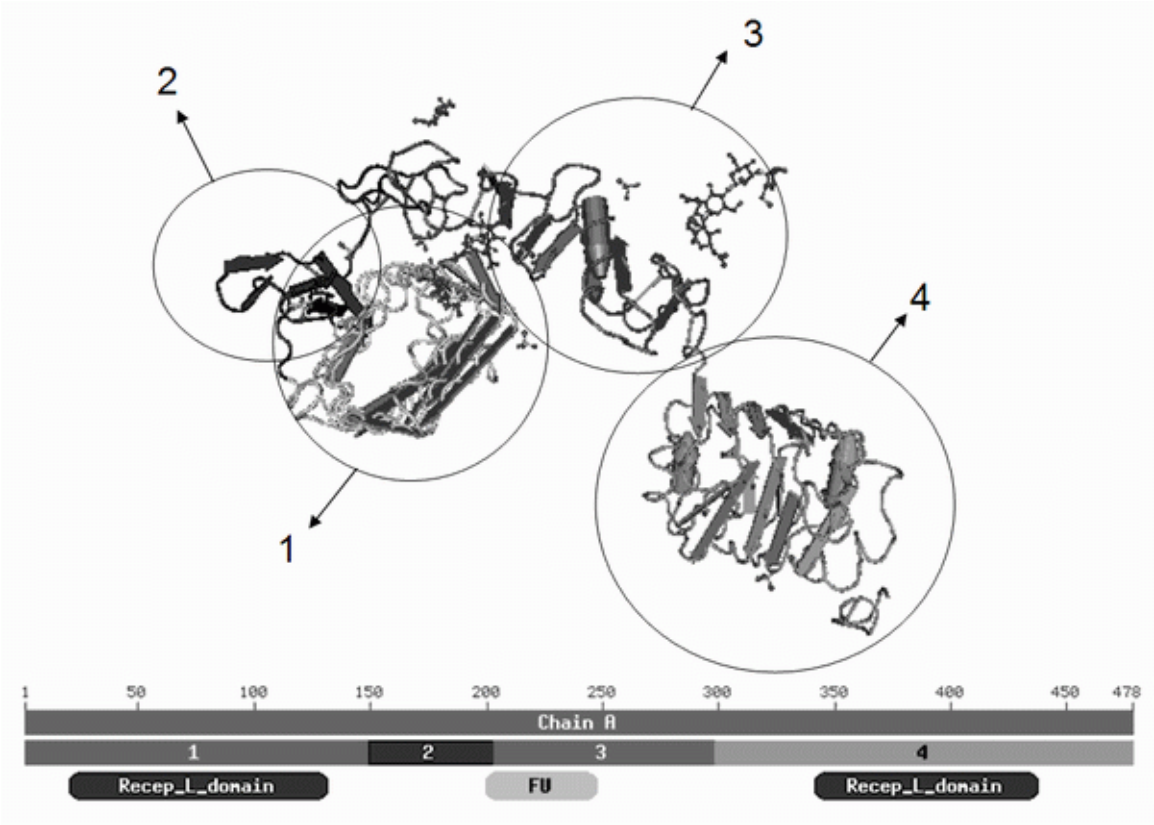


Figura 4 – Exemplo de domínios de uma proteína[6]

Mutações são, simplesmente, alterações na seqüência de DNA. Modificações são introduzidas na seqüência de DNA de um organismo devido a razões diversas, tais como a incidência de radiação ou a atuação de agentes químicos. Mutações sucessivas causam alterações nas seqüências de DNA de modo que sua função se altera ou deixa de ser reconhecida (Bedell et al, 2003). O próprio processo de replicação de DNA das células tem imperfeições e a taxa de erro do processo de replicação de DNA é dito ser de 1 erro a cada 300 milhões de bases. Uma vez que o genoma humano, por exemplo, tem 3 bilhões de bases, são esperadas cerca de 10 mutações a cada replicação. Considerando que um ser humano possui cerca de 1 trilhão de células, espera-se que inúmeras mutações ocorram.

Entretanto, mutações significativas são as que ocorrem em regiões do DNA que codificam proteínas. Uma vez que o DNA é alterado, o mesmo ocorre com o mRNA, o

[6] Figura retirada de http://www.ncbi.nlm.nih.gov/

que pode então levar a uma mudança na seqüência da proteína produzida. No entanto, isso não ocorre todas as vezes, devido à redundância do código genético. Há mutações potencialmente mais prejudiciais, que alteram um códon de maneira a criar um códon de terminação (referidos como “Ter” na figura do código genético), e, assim, fazem com que a tradução do mRNA seja terminada precocemente.

Outro tipo de mutação são as que inserem ou removem nucleotídeos. Além delas, existem duplicações, inversões e outros rearranjos de genes que podem uni-los ou destruí-los. Mutações que inserem ou removem nucleotídeos também são freqüentemente destrutivas, pois quando um número de nucleotídeos que não é um múltiplo de 3 é inserido ou removido das seqüências de DNA, altera-se a fase de leitura, causando diversas diferenças nos códons lidos dali em diante.

Assim, séries de mutações, associadas a outros fatores evolutivos, acabam por criar novas espécies, e internamente isso representa a alteração de diversos genes num organismo. Genes ou seqüências derivadas de um ancestral comum são chamados de homólogos. Homólogos podem ainda ser divididos entre parálogos e ortólogos. Dizem-se ortólogos os criados a partir de eventos de especiação, ou seja, seriam “o mesmo gene”, mas encontrado em espécies diferentes. Já parálogos são criados por eventos de duplicação gênica. Logo, dois genes duplicados num mesmo organismo são chamados parálogos.

Boa parte dos processos de identificação de seqüências é feita baseado na procura de homologias, e para isso costumam ser usadas buscas de similaridade. Entretanto é importante distinguir os dois termos, pois ainda que uma seqüência apresente similaridade com outra seqüência em, por exemplo, 40% de sua extensão, não se pode dizer que elas são 40% homólogas. Análises devem ser feitas de modo a inferir com certeza se elas são ou não homólogas, baseado tanto em sua similaridade quanto em outros fatores.

3. Anotação de Seqüências

O resultado do seqüenciamento genético é simplesmente uma seqüência de aminoácidos ou nucleotídeos sem qualquer identificação. Assim, a anotação é um processo que tem por objetivo agregar informação biológica a essas seqüências genéticas por meio de sua identificação funcional. Ou seja, mediante diversas análises, a anotação visa a descrever biologicamente o que foi seqüenciado. “A priori”, a anotação pode ser dividida em três etapas, como detalhado abaixo: no nível de nucleotídeo, no nível protéico e a anotação no nível de processos (Claverie et al, 1997; Eisenberg et al, 2000; Grayhack e Phizicky, 2001, Stein, 2001).

3.1. Níveis de anotação

A fase inicial da anotação, feita no nível de nucleotídeos tem como atividade principal a procura de genes na seqüência de DNA e a localização de marcadores por meio de mapeamentos realizados por análises biológicas prévias. Este conjunto de marcadores funcionam então como pontos de referência para a análise subseqüente: a procura por genes. Uma vez identificados os genes, são então identificadas seqüências correspondentes a RNAs não codificadores, seqüências regulatórias, elementos repetitivos e polimorfismos (Stein, 2001).

Após a anotação no nível de nucleotídeos, inicia-se a etapa de anotação no nível protéico. Esta etapa é constituída pela nomeação das proteínas do organismo e pela associação de possíveis funções a estas proteínas. Neste caso, são utilizados bancos de dados de seqüências primárias, estruturais, de famílias gênicas ou de domínios funcionais como as bases SWISSPROT[7], Protein Data Bank[8] (PDB) ou CDD[9] (Doerks et al., 1998; Whisstock e Lesk, 2003; Marchler-Bauer, et al., 2005).

[7] SWISSPROT - http://www.ebi.ac.uk/swissprot
[8] Protein Data Bank - http://www.rcsb.org/pdb
[9] CDD - http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

Depois destes dois níveis tem início então a etapa de anotação no nível de processos. Esta etapa tem como objetivo relacionar o genoma a processos biológicos, isto é, estabelecer como os constituintes de um genoma se relacionam com o ciclo celular, a morte celular, embriogênese, metabolismo e manutenção da saúde do organismo. Esse processo depende da existência de um banco de dados dotado de um esquema de classificação associado a funções biológicas conhecidamente descritas, com especificidade suficiente para distinguir entre proteínas que sejam membros de uma mesma família gênica. A base de dados Gene Ontology (GO) criada em 1991 é um repositório dessa natureza (Ashburner et al., 2000). O GO é constituído de um vocabulário padrão que descreve funções de genes eucarióticos. Essa base é subdividida em três grupos: funções moleculares, que descrevem tarefas realizadas por produtos gênicos individuais; processos biológicos, que descrevem processos de maneira ampla como meiose, por exemplo, e componentes celulares que associam produtos gênicos a estruturas subcelulares como organelas (Faria-Campos, 2005). Também são bases desse tipo o COG/KOG (Tatusov et al., 2000), e as bases CGAP-KEGG[10] (Kanehisa et al, 2004) e CGAP-BioCarta[11].

3.2. Procura de genes

Os diversos níveis de anotação são igualmente relevantes e estão interligados. Contudo, a procura e identificação dos genes é a etapa central da anotação. Em genomas de procariotos essa etapa é realizada sem maiores dificuldades, uma vez que ela consiste basicamente na identificação de janelas abertas de leitura (em inglês “Open Reading Frames” ou ORFs) na seqüência produzida. Em eucariotos, por outro lado, o processo de busca de genes é complicado pela presença de íntrons e sítios de “splicing” alternativo, que representam descontinuidades da seqüência de um gene dentro do genoma. Por essa razão, métodos diversos para a predição de genes em

[10] CGAP-KEGG – http://cgap.nci.nih.gov/Pathways/Kegg_Standard_Pathways
[11] CGAP-BioCarta – http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways

seqüências eucarióticas têm sido amplamente utilizados (Cho e Walbot, 2001; Misra et al., 2002).

De maneira geral, a procura por genes é feita a partir de dois tipos de métodos de predição distintos, designados intrínsecos e extrínsecos. Os métodos para predição intrínsecos são baseados no reconhecimento de características específicas do gene em associação com a análise do conteúdo da seqüência. Características específicas da seqüência normalmente associadas à presença de genes (promotores, códons de início e finalizadores, sítios de splicing, etc.) são utilizadas como sinais para inferir a presença de um gene, juntamente com a diferença da distribuição de nucleotídeos apresentada entre regiões que contêm genes e regiões intergênicas. A combinação da informação proveniente destes padrões permite não só a localização de genes completos em uma seqüência genômica como também de estruturas gênicas parciais nas extremidades da seqüência analisada (Gibas e Jambeck, 2001). Entretanto esses métodos costumam ser específicos para determinado organismo (ou grupo de organismos) já que o modo como os genes são codificados no DNA varia consideravelmente entre tipos de organismos. Além disso, mesmo quando é possível encontrar um gene com um método intrínseco, não se obtém nenhuma informação relativa à função biológica desse gene. No entanto, estima-se que cerca de 30% a 50% dos genes de um genoma não possuem homólogos conhecidos (valor que está diminuindo à medida que mais genes são identificados) ou apresentam similaridade pequena (menos de 30% de identidade) com proteínas conhecidas, o que torna impossível a anotação confiável dessas seqüências utilizando buscas de similaridade (Yakunin et al., 2004). Assim, a simples descoberta da existência do gene, mesmo que dela não provenha mais informações, já é importante para permitir que a pesquisa continue (Faria-Campos, 2005).

Por outro lado, em genomas recém-seqüenciados genes são anotados primariamente com base em sua homologia com proteínas já caracterizadas em outros genomas. Esse enfoque é designado extrínseco por desconsiderar as características existentes na seqüência investigada. Nesse tipo de abordagem são utilizados os programas baseados em busca de similaridade que têm como premissa o fato de que a

conservação existente entre as seqüências de diferentes espécies implica também a conservação funcional. Tais programas buscam a similaridade existente entre uma região genômica desconhecida ou uma seqüência não identificada e uma seqüência de proteína ou nucleotídeos presente em um banco de dados, para determinar então se a região em questão é ou não uma região codificadora (e toda e qualquer informação a mais que esteja disponível). A busca de similaridade é feita mediante alinhamentos entre seqüências e a subseqüente classificação desses alinhamentos. Os alinhamentos entre duas seqüências podem ser globais, nos quais duas seqüências são alinhadas ao longo de toda a sua extensão, ou locais em que segmentos parciais de duas seqüências são alinhados até a máxima extensão possível. Os alinhamentos globais assumem que as duas seqüências a serem alinhadas compartilham regiões similares, e as alinham em toda a sua extensão. As ferramentas de alinhamento global utilizam buscas exaustivas e são comparativamente menos eficientes do que as ferramentas para alinhamentos locais. Assim, quando seqüências desconhecidas são utilizadas, os alinhamentos locais são mais apropriados, sendo utilizados quando se quer anotar uma seqüência-teste a partir de uma seqüência homóloga a ela presente numa grande base de dados (Baxevanis e Oullette, 2001; Gibas e Jambeck, 2001).

Na ferramenta descrita neste trabalho, utiliza-se apenas métodos extrínsecos de procura e identificação de genes.

3.3. Identificação dos genes e buscas de similaridade

Nos diversos métodos ditos extrínsecos para a identificação de genes, em geral o primeiro passo é a comparação da seqüência-teste (“query”) contra uma base de seqüências previamente identificadas. Por meio dessa busca, pode-se, então, encontrar seqüências que apresentam um nível significativo de similaridade com a seqüência “query” em questão. Para essa busca, a técnica mais eficiente é a busca de alinhamentos locais e a subseqüente análise dos resultados (“hits”) encontrados.

Atualmente a ferramenta mais utilizada nas buscas de homologia que fazem uso de alinhamentos locais é o pacote BLAST (Basic Local Alignment Search Tool – Altschul et al., 1990). O BLAST compara uma ou mais seqüências (“query”) em um arquivo texto a uma base de dados previamente determinada e formatada (seqüências “subject”). As seqüências “query” devem estar num arquivo texto em algum dos diversos formatos aceitos pelo BLAST (FASTA, ASN.1, etc). O mais comum desses formatos, que será usado nesse trabalho, é o FASTA. Já as seqüências “subject”, contra as quais as seqüências “query” serão comparadas, devem ser formatadas, usando para isso um programa do próprio pacote BLAST (“formatdb”), numa base em que o BLAST irá consultar.

Um arquivo de texto no formato FASTA pode conter uma ou mais seqüências, de forma que cada seqüência é composta por uma linha de identificação e as linhas subseqüentes contêm a informação da seqüência propriamente dita. Para as linhas contendo a seqüência, recomenda-se que elas tenham menos de 80 caracteres (em geral se usam 70 caracteres em cada linha). A linha de identificação começa com um “>” e em seguida, sem espaços, vem a identificação da seqüência. Em geral, logo depois do “>” há os identificadores da seqüência separados por “|” e em seguida outras informações como a descrição do que é a seqüência ali representada e às vezes a origem da seqüência, organismo a que pertence, quem a seqüenciou ou onde foi produzida. Abaixo um exemplo de uma seqüência no formato FASTA.

```
>gi|556413|gb|AAA93516.1| phosphoglycerate kinase [Schistosoma mansoni]
MGLSKLSISDVDLKGKRVLIRVDFNVPMKDGKVTNTQRIAAAIPTIKYALDKGAKSVVLMSHLGRPDGHK
VDKYSLKPVCPEVSKLLGKEVTFLNDCVGPDVVNACANPAPGSVFLLENLRFHVEEEGKGVSPTGEKTKA
TADQIKAFSESLTKLGDVYVNDAFGTAHRAHASMVGCQLPQKACGFLMNKELTYFAKALENPERPFLAIL
GGAKVSDKIQLINNMLDKVNELIIGGGMAYTFLKQIHNMHIGNSLFDAPGAEIVHKVMETAKAKNVAIHL
PVDFVTADKFADDANTEIRTIQSGIADGWMGLDIGPKTIEEFSKVISRAKTIVWNGPMGVFEMDKFATGT
KAAMDEVVKATKNGATTIIGGGDTATCCAKWDTEDKVSHVSTGGGASLELLEGKQLPGVVALTDAH
```

Assim, uma pesquisa BLAST gera um relatório no qual são listados todos os alinhamentos entre duas seqüências que atendem aos parâmetros especificados, juntamente com os valores de “score” (às vezes referido como “s”) e as estatísticas

avaliando a probabilidade de ocorrência ao acaso do alinhamento entre as duas seqüências (“Expected value”, ou “e”). Os valores de “s” e “e” são utilizados para escolher o melhor alinhamento determinado para uma dada seqüência. Quanto maior o valor de “s” ou menor o valor de “e”, melhor o alinhamento (Baxevanis e Oullette, 2001). O algoritmo utilizado pelo BLAST permite que as comparações entre seqüências sejam feitas de maneira extremamente eficiente, e isso tornou esse pacote o padrão como método de busca de homologia entre seqüências e o programa mais utilizado pela maioria dos serviços atualmente encontrados (Altschul et al., 1997; Gibas e Jambeck, 2001). O BLAST se encontra disponível para uso remoto em diversos servidores, por exemplo, no “website” do National Center for Biotechnology Information (NCBI)[12], ou pode ser utilizado localmente por meio da instalação do binário disponibilizado também por esta instituição. A utilização do BLAST localmente exige, além da instalação, a construção de bases de dados locais a serem utilizadas como “subject” nas buscas.

3.4. BLAST e programas de alinhamento local

O principal método usado atualmente para busca de homologias é o de busca de alinhamentos locais. Para isso existem hoje diversos programas que realizam essa tarefa, sendo que o BLAST é o mais usado.

Os programas que fazem busca de alinhamentos locais implementam variações do algoritmo chamado “Smith-Waterman”, que tem seu nome herdado de seus criadores. O algoritmo, que utiliza uma abordagem de programação dinâmica, fornece uma solução ótima, indicando os melhores alinhamentos locais para o sistema de pontuação usado (Smith e Waterman, 1981).

Entretanto, na prática o tamanho expressivo dos dados faz com que o algoritmo “Smith-Waterman” não seja usado em favor de heurísticas mais eficientes. Dentre essas abordagens heurísticas, a mais usada é a implementada no pacote de programas

[12] National Center for Biotechnology Information - http://www.ncbi.nlm.nih.gov/

BLAST, que consegue identificar a grande maioria dos alinhamentos que atendem o critério desejado com um ganho de performance significativo (Altschul et al, 1990 e Altschul et al, 1997).

O BLAST possui programas para alinhamentos de seqüências de proteínas e nucleotídeos, além de outros para criação de bancos de dados que serão então usados nas comparações. Numa pesquisa BLAST tem-se um arquivo contendo as seqüências para as quais se deseja encontrar alinhamentos (“query”) e um banco de dados de seqüências com as quais se vai comparar (“subject”).

Dependendo do tipo de seqüências de cada grupo, usa-se um programa diferente:

BLASTN – Nucleotídeo / Nucleotídeo
BLASTP – Proteína / Proteína
BLASTX – Proteína / Nucleotídeo traduzido em proteína
TBLASTN – Nucleotídeo traduzido em proteína / Proteína
TBLASTX – Nucleotídeo traduzido em proteína / Nucleotídeo traduzido em proteína

Ao contrário do algoritmo Smith-Waterman que garante encontrar todos os alinhamentos locais que atendam a uma certa restrição de pontuação mediante uma abordagem de programação dinâmica, o BLAST adota uma abordagem heurística. O algoritmo do BLAST é realizado em 3 passos (Figura 4):

a) Criação de uma lista com todas as “palavras” de tamanho “w” para as seqüências da base subject, que então são comparadas a palavras de mesmo tamanho na “query”, de modo a encontrar as que resultem num “score” de no mínimo “T”

b) Para cada palavra que atenda à essa restrição, então se encontra o alinhamento exato (“hit”) na seqüência da base de dados

c) Estendem-se então os alinhamentos enquanto o “score” se mantém acima do limite aceitável

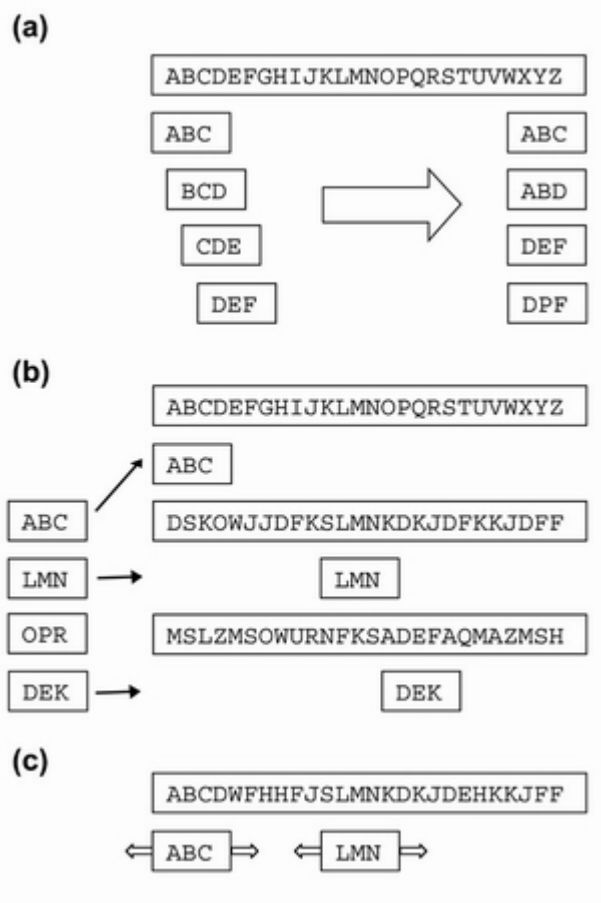


Figura 5 – Algoritmo do BLAST

O cálculo do “score” para alinhamentos de seqüências de aminoácidos é feito baseado numa matriz de substituição, que leva em conta as características dos aminoácidos para determinar a pontuação quando aminoácidos diferentes são alinhados. Para aminoácidos iguais e “gaps” inseridos no alinhamento há uma pontuação fixa.

Para alinhamentos de seqüências de nucleotídeos em geral não se usa uma dessas matrizes de substituição. Tem-se uma pontuação fixa, porém distinta, para cada

uma das situações possíveis: um acerto, um erro ou um “gap” no alinhamento. Sendo que quando há um erro, o “score” é sempre o mesmo.

Por fim, os alinhamentos são avaliados sob o ponto de vista estatístico, no sentido de identificar a probabilidade de eles terem ocorrido ao acaso, que é calculado com base no tamanho da “query”, da base, do alinhamento e do tipo de seqüência alinhada. Esse resultado é expresso no valor de “e”, chamado de “Expect” e o BLAST permite que só se considerem alinhamentos com valor de “e” numa determinada faixa.

Com isso, cada alinhamento que atende aos requisitos de “score” e “expect” é relatado junto com os respectivos valores de “s” e “e”.

Ainda que o BLAST adote uma abordagem heurística em comparação com o algoritmo ótimo de programação dinâmica Smith-Waterman, ambos tem complexidade temporal O(mn), sendo “m” o tamanho da seqüência “query” e “n” o tamanho da base. Ou seja, o tempo que se gasta para executar ambos algoritmos varia proporcionalmente com produto dos tamanhos da “query” e da base a ser pesquisada. Entretanto, a implementação mais recente do BLAST chega a ser mais de 100 vezes mais rápida do que implementações do algoritmo Smith-Waterman, sendo que o BLAST só deixa de encontrar menos de 0,6% dos alinhamentos que o Smith-Waterman detecta (Altschul et al, 1997). Uma troca de sensibilidade por eficiência que se mostra bastante compensadora, dado que o tamanho dos dados não só é muito grande como eles tendem a crescer mais rápido do que a capacidade de processamento dos computadores atuais.

3.5. Análise de estrutura de domínios

Para obter uma anotação precisa de uma seqüência deve-se buscar outros meios para validar os resultados encontrados pelo BLAST. A comparação da estrutura

de domínios da seqüência “query” com a estrutura do “hit” reportado pelo BLAST permite que se aumente a confiabilidade da anotação.

Os domínios de uma proteína fornecem importante informação sobre sua função biológica, e a comparação dos domínios encontrados em duas seqüências permite ao usuário confirmar se duas seqüências apresentam ou não homologia, e ainda pode ser útil no sentido de permitir selecionar qual “hit” mais se relaciona à seqüência “query”.

3.6. Análise filogenética

Outro meio para validar o resultado obtido com o BLAST e, conseqüentemente, obter uma anotação mais confiável, é a análise filogenética. Métodos de análise filogenética utilizam alinhamentos globais entre múltiplas seqüências e buscam determinar uma relação evolutiva entre elas.

Existem vários métodos para a geração de árvores filogenéticas e entre eles se destacam os de parsimônia, matriz de distâncias e de similaridade máxima. Uma vez que se tem uma árvore filogenética que contém a seqüência “query” e o “hits” obtidos do BLAST, pode-se observar quais as seqüências mais próximas, evolutivamente, da seqüência “query” e em seguida validar os resultados previamente obtidos

4. Trabalhos Relacionados

A anotação de seqüências usando bases secundárias ganhou significativa importância nos últimos tempos pela maior confiabilidade de seus dados e da possibilidade de se determinar a função biológica e grupo funcional de uma seqüência com relativa facilidade. Com isso, diversas bases desse tipo foram criadas e a maioria delas disponibiliza, juntamente com as seqüências e sua informação associada, “front ends” para execução do BLAST contra essas seqüências, de modo que seria possível classificar uma seqüência-teste segundo a categorização feita pela base.

Outras ferramentas, assim como a PCT descrita nesse trabalho, utilizam-se dessas bases para promover a anotação de seqüências às vezes somando à identificação provida pelas bases secundárias outros métodos de anotação.

Abaixo segue uma análise de diversos desses serviços, separados em bases secundárias e ferramentas de anotação.

4.1. Bases Secundárias

Bases primárias funcionam como depósitos de informação biológica (por exemplo, seqüências), que em geral se trata de resultados de experimentos com alguma interpretação, mas sem uma revisão profunda. Um exemplo de base primária de seqüências é o NR do GenBank.

Por outro lado, as bases secundárias contêm informação derivada das bases primárias que passam por uma curadoria. Essas bases têm o propósito de servir como bancos de dados de seqüências devidamente curadas, e que além da classificação individual provêm também uma classificação funcional, agrupando as seqüências

segundo algum tipo de classificação biológica. Abaixo segue uma descrição de algumas dessas bases.

4.1.1. UniProt/GOA

O projeto “Gene Onthology Annotation” (GOA) desenvolvido pelo “European Bioinformatics Institute” (EBI) tem como objetivo a produção de um vocabulário informativo de genes que pode ser atribuído a todos os organismos eucarióticos. Na base GOA, os genes não estão organizados em vias bioquímicas, mas numa rede hierárquica de termos que descrevem os atributos dos produtos gênicos. Dessa forma, uma entrada pode possuir vários identificadores GOA agregados e a cada avanço nos conhecimentos a seu respeito, novos identificadores podem ser adicionados. Atualmente esta base contém 2.388.845 entradas, curadas pelo consórcio UniProt e com termos de ontologia atribuídos pelo GOA (Harris et al., 2004).

Com a união de esforços entre os grupos responsáveis por PIR, Swiss-Prot e TrEMBL, surgiu a base UniProt[13] (Apweiler et al., 2004). Esse consórcio tem como objetivo compreender, em uma base única, todas as proteínas seqüenciadas até o momento no mundo. Além disso, existe a preocupação constante com a anotação funcional das seqüências depositadas, resultando em uma base pública rica, coerente e com posicionamento biológico.

No site do UniProt, há uma interface para realização de comparações de similaridade usando BLAST contra a sua base de seqüências. Para os hits encontrados, é possível ver extensas informações relativas à seqüência, como dados de ontologia fornecidos pelo GOA e referências bibliográficas que tratam da seqüência (Figura 5).

[13] UniProt - http://www.uniprot.org

Todavia, a busca, que é feita online, demora consideravelmente. A ferramenta está disponível em http://www.pir.uniprot.org/search/blast.shtml

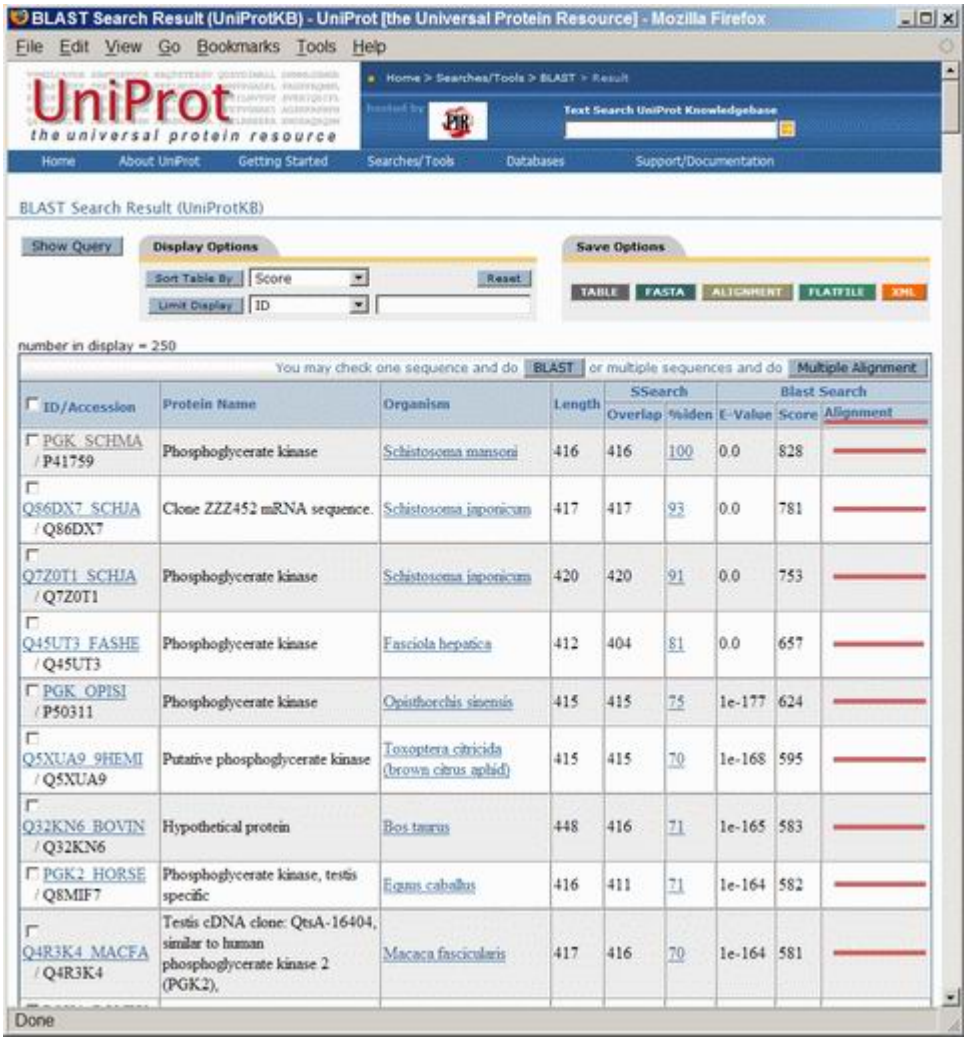


Figura 6 – Tela de resultados do UniProt/GOA

4.1.2. COG / KOG

Organizada pelo NCBI, a base COG[14] ("Cluster of Orthologous Groups") representa um agrupamento de proteínas ortólogas e foi produzida por meio de comparações entre seqüências de quarenta e três genomas de organismos procarióticos. Cada um desses agrupamentos (nesse caso, chamados individualmente de COG) corresponde a uma entrada composta por proteínas distintas ou grupos de parálogos presentes em ao menos três linhagens, correspondendo a domínios evolutivamente conservados (Tatusov et al., 2000). Assim, cada COG representa um conjunto de genes e seus ortólogos, os quais possuem a mesma função biológica, sendo estes, por sua vez, agrupados em categorias funcionais. Uma versão mais ampla da base COG foi criada, expandindo a lista de genomas, para incluir seqüências de organismos eucariotos. Ela foi intitulada KOG[15], numa alusão à adição de seqüências eucarióticas (eu**k**aryotic) na atualização da versão anterior (COG) e também contém classificações das proteínas em categorias funcionais (Tatusov et al, 2001). Para figurar nessa variante, a entrada deve seguir as mesmas exigências da versão anterior – os ortólogos devem estar presentes em pelo menos três organismos.

Atualmente a base COG contém 144.320 seqüências, distribuídas em 3.280 COGs, e a extensão KOG contém 88.654 seqüências, distribuídas em 4.607 entradas. Assim, quando é identificada homologia em relação a proteínas dessas bases, é possível propagar também a classificação funcional simultaneamente à anotação.

A página do projeto COG (Tatusov et al, 2003.) disponibiliza uma ferramenta que permite que seja feita uma comparação de similaridade usando o BLAST contra as seqüências do projeto, em que para os hits é atribuída a classificação provida pelo COG (identificação da proteína e a categoria funcional a que ela pertence). A interface não dá muitas opções ao usuário e só funciona para seqüências de aminoácidos.

[14] Cluster of Orthologous Groups - http://www.ncbi.nlm.nih.gov/COG/
[15] Clusters of euKaryotic Orthologous Groups - http://www.ncbi.nlm.nih.gov/COG/grace/shokog.cgi

Existem duas versões do aplicativo, uma para cada versão da base COG: http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html e http://www.ncbi.nlm.nih.gov/COG/grace/kognitor.html (Figura 6).

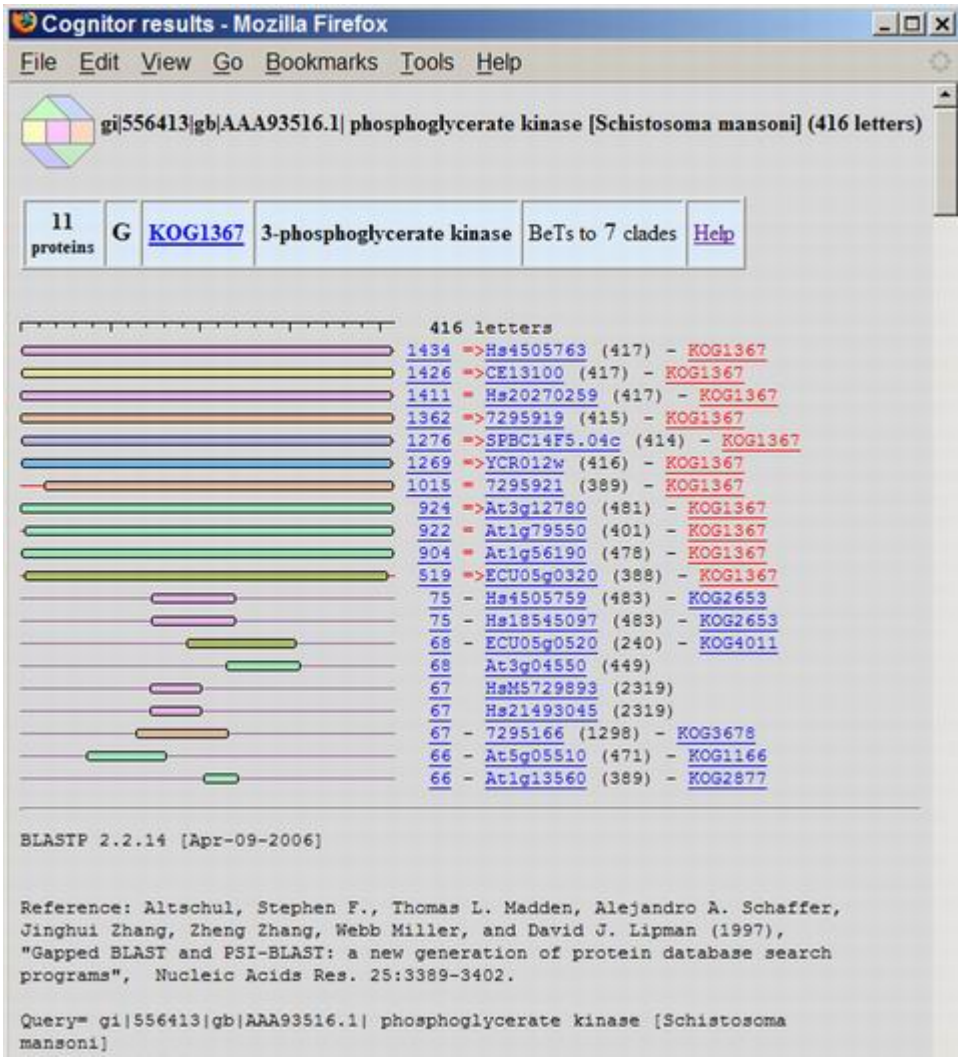


Figura 7 – Tela de resultados do KOGnitor

4.1.3. KEG

A “Kyoto Encyclopedia of Genes and Genomes” (KEGG) é um projeto que visa criar uma base de conhecimento de informações genéticas, ligando funções de genes conhecidos com informações funcionais de mais alto nível (Kanehisa, M., et al., 2004).

O projeto foi iniciado pelo junto ao programa de genoma humano japonês e provê um serviço de comparação de seqüências entradas pelo usuário contra as seqüências do projeto. Quando é encontrado algum hit, é fornecida informação relativa e ele, indicando a classificação e sua respectiva via bioquímica. Esse serviço funciona "online" e aceita tanto seqüências de nucleotídeos quanto de aminoácidos. O serviço está disponível em: http://www.genome.jp/kegg-bin/kaas_main (Figura 7).

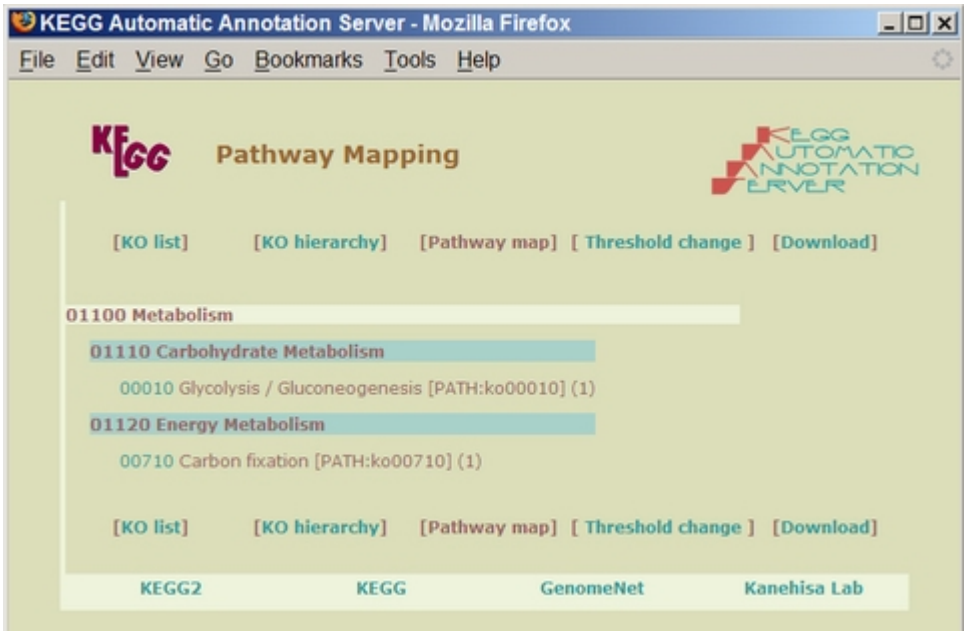


Figura 8 – Tela de resultados da ferramenta de anotação do projeto KEGG

4.1.4. CGAP

O "Cancer Genome Anatomy Project" (CGAP) consiste num programa interdisciplinar com o objetivo de gerar informações e ferramentas necessárias para o estudo da anatomia molecular da célula do câncer. O programa é administrado pelo "National Cancer Institute" dos Estados Unidos e tem como colaborador o NCBI.

O CGAP organiza genes e proteínas catalogados pelo projeto em vias bioquímicas utilizando informação provida pelo projeto KEGG e pela empresa

BioCarta[16]. Através de sua página, o CGAP provê informação sobre os genes e proteínas com gráficos e figuras, além de permitir o “download” das seqüências, mas não há nenhuma funcionalidade de busca ou comparação de seqüências.

4.1.5. CDD

O CDD (Conserved Domain Database) é uma base de domínios de proteínas organizada pelo NCBI, cujos domínios são reunidos a partir de outras bases, sendo que as principais são o SMART[17], o Pfam[18] e o COG. Ainda que a o CDD tenha sido formado a partir do conteúdo dessas outras bases, ele tem se desenvolvido independentemente, sendo atualizada com dados provenientes de curadorias desenvolvidas pelo próprio NCBI.

O CDD é usado como componente de classificação de proteínas do sistema Entrez[19] do NCBI.

4.2. Ferramentas de anotação

4.2.1. NCBI BLAST

Na página do BLAST no NCBI é possível fazer pesquisas BLAST contra a base NR ou ainda contra algumas outras bases menores. A página inclui várias opções de uso dependendo do tipo de seqüência usado ou da base contra a qual se vai comparar, mas deve-se usar uma página específica, o que pode criar alguma confusão. O serviço ainda permite a exibição de uma árvore filogenética dos resultados e também aponta “hits” de domínios de proteínas para a seqüência utilizada.

[16] Biocarta - http://www.biocarta.com/
[17] SMART - http://smart.embl-heidelberg.de/
[18] Pfam - http://www.sanger.ac.uk/Software/Pfam/
[19] Entrez - http://www.ncbi.nlm.nih.gov/sites/gquery

4.2.2. Blast2GO

A ferramenta Blast2GO (Conesa et al, 2005) é implementada em Java e possui versões tanto “online” quanto para “download”. Na seqüência de uso do Blast2GO, o primeiro passo é carregar um arquivo contendo as seqüências a serem anotadas e em seguida realizar uma pesquisa BLAST. O usuário deve selecionar a base de dados contra qual será realizada a busca (NR ou Swissprot) e o programa do BLAST a ser usado (blastp, blastn, etc.) e então o programa se conecta a um servidor e realiza a pesquisa BLAST. Em seguida, o usuário pode fazer um mapeamento entre os “hits” encontrados e as classes de ontologia do GOA.

Por fim, pode-se visualizar gráficos das classes de ontologia e estatísticas sobre a anotação das seqüências. A Figura 8 mostra uma tela de resultados do Blast2GO.

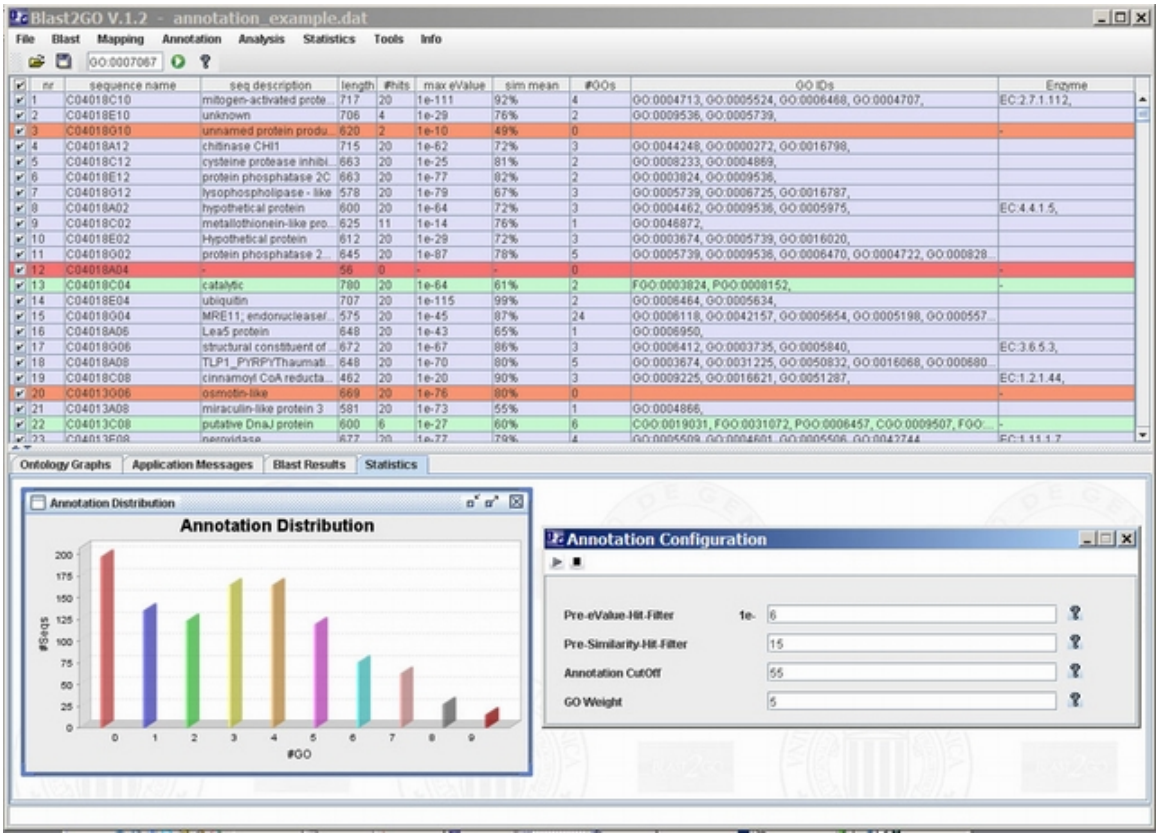


Figura 9 – Tela de resultados da ferramenta Blast2GO

4.2.3. AutoFACT

A “AutoFACT” (Koski et al, 2005) é uma ferramenta implementada em Perl com uma versão “online” e outra disponível pra download. Essa ferramenta permite que o usuário faça uma pesquisa BLAST de suas seqüências contra algumas bases secundárias (COG, KEGG e UniRef) e o NR alem de algumas bases de domínios (Pfam e Smart). Uma vez que são encontrados “hits” nas buscas contra essas bases, seria possível atribuir informação de qualidade à seqüência teste utilizando tais “hits”, informando inclusive a via bioquímica ou categoria funcional a que a seqüência faria parte, de acordo com a classificação provida pela base em questão. Entretanto a versão online dessa ferramenta não está mais acessível e não foi possível instalar a versão para download.

4.2.4. GARSA

“GARSA” (Davila et al, 2005) é uma ferramenta para integração de informação biológica. A ferramenta é implementada usando Perl, CGI, Apache e MySQL para funcionar via web. Dentre as funcionalidades relatadas está a capacidade de usar como entrada cromatogramas, arquivos fasta locais ou retirados do GenBank e a capacidade de analisar esses dados usando comparações BLAST e análises filogenéticas. Porém, o “site” onde a ferramenta[20] está hospedada não funciona corretamente (“login” não funciona) e o contato realizado para fazer o “download” da ferramenta não teve resposta, impedindo uma análise mais profunda de suas funcionalidades.

4.2.5. SABIA

O “SABIA” é uma ferramenta para montagem e anotação de genomas de organismos procariotos (bactérias). A ferramenta realiza tarefas de montagem

[20] “Site” do GARSA: http://garsa.biowebdb.org/

automática, detecção de regiões codificadoras e análise de regiões extragênicas. A ferramenta integra vários softwares de análise e algumas bases secundárias. A ferramenta está disponível para download a partir do site do projeto mediante requisição.

4.2.6. Comparação entre ferramentas de anotação

A Tabela 3 abaixo sumariza a comparação entre as ferramentas analisadas nesse trabalho. Nela vemos que a PCT se destaca por implementar todas as funcionalidades de interesse desse trabalho. Algumas ferramentas como o SABIA e o GARSA apresentam outras funcionalidades úteis na análise de cromatogramas e de genomas completos, que não estão no escopo desse trabalho.

Ferramenta	**Modo de uso**	**Uso de bases secundárias**	**Análise filogenética**	**Análise de domínios**
NCBI BLAST	Online e local	Não, apenas bases primárias	Sim	Sim
Blast2GO	Online e local	Sim, apenas GOA	Não	Não
AutoFACT	Online e local	Sim várias	Não	Sim
GARSA	Online e local	Não.	Não	Não
SABIA	Local	Sim, COG e GOA	Não	Não
PCT	Online e local	Sim, várias	Sim	Sim

Tabela 3 – Ferramentas de anotação

Todas elas permitem o funcionamento local, entretanto o processo de instalação dessas ferramentas nem sempre é trivial ou funciona como anunciado. Esse é outro ponto no qual a PCT se destaca por ser de fácil instalação. Sendo uma ferramenta que funciona via web, sua instalação consiste na cópia dos arquivos, ajuste de permissões de acesso, importação do banco de dados MySQL e edição de um arquivo de configuração para ajustar nome de usuário e senha do banco.

5. PCT: Funcionamento e Implementação

A PCT propõe então um método de anotação que tem como base a comparação de similaridade contra bases secundárias usando o BLAST. Quando são encontrados hits, eles podem ser analisados por meio de outras duas funcionalidades:

- Alinhamento global e então a geração de uma árvore filogenética, usando o pacote Phylip (Felsenstein, J., 2005)

- Comparação dos domínios conservados dos best hits de cada base

Ainda é possível fazer uma pesquisa BLAST da seqüência query contra a base NR caso não seja encontrado nenhum hit, ou caso se deseje fazer uma comparação dos hits encontrados com os hits de uma base significativamente maior como a NR.

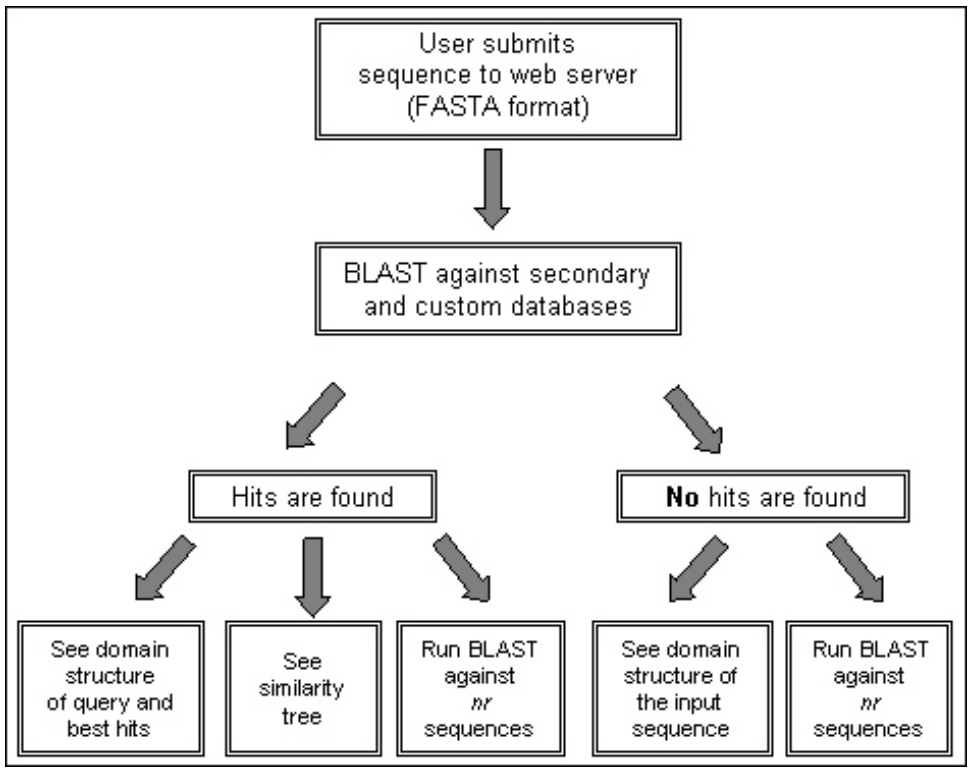


Figura 10 – Estrutura de funcionamento da PCT

A ferramenta foi implementada em PHP (originalmente em PHP4, mas funciona em PHP5) e utiliza MySQL versão 5.0.27. As comparações BLAST realizadas são feitas usando o pacote BLAST versão 2.2.13 obtido do NCBI. A ferramenta ainda faz uso de um script Perl para formatação de seqüências a serem usadas na geração da árvore.

A implementação da ferramenta está disponível para uso online via web no endereço http://biotec.icb.ufmg.br/pct2 e para "download" em http://biotec.icb.ufmg.br/pct2/help.html.

5.1. As bases de dados

O núcleo da PCT é a funcionalidade de BLAST contra bases secundárias. Para cada base, tem-se uma coleção de seqüências associadas a uma tabela MySQL que contém informação relativa a cada uma dessas seqüências. Essa estrutura permite, uma vez encontrado um "hit" com a seqüência do usuário, que se retorne a identificação desta seqüência (o nome da proteína) e a informação associada que indica a via bioquímica ou categoria funcional a que pertence essa seqüência.

Quando um "hit" é encontrado, o identificador da seqüência é usado para que se encontre no banco de dados a informação associada à seqüência em questão, e em seguida essa informação é mostrada ao usuário.

A estrutura básica comum das tabelas de cada base é a seguinte:

```
| id        | classification   | pathway                        |
+-----------+------------------+--------------------------------+
| NM_000668 | Alcohol dehy...  | Glycolysis_Gluconeogenesis     |
| NM_000668 | Alcohol dehy...  | Fatty Acid Metabolism          |
| NM_000668 | Alcohol dehy...  | Bile Acid Biosynthesis         |
```

Onde “id” é o identificador da seqüência, “classification” é o nome do gene ou proteína em questão e “pathway” provê uma classificação funcional ou da via bioquímica a que a seqüência pertence.

A estrutura modular da ferramenta permite ainda que outros campos sejam acrescentados a cada uma dessas tabelas e então tratados e exibidos conforme desejado.

As bases secundárias de seqüências descritas na seção anterior foram obtidas de seus respectivos websites e as seqüências fasta são usadas para o BLAST inicial e as informações associadas (armazenadas em tabelas no MySQL como descrito acima) usadas para identificação dos “hits” encontrados. A Tabela 4 abaixo sumariza a informação relativa às bases utilizadas na PCT.

Nome da base	Total de seqüências	Classes	Proteínas distintas
CGAP Biocarta	11,177	313	2,039
CGAP Kegg	2,877	95	1,136
COG	144,320	143	3,280
KOG	88,645	228	4,607
GOA	6,584,517	8,798	381,756

Tabela 4 – Composição das bases secundárias

5.2. Análise de estrutura de domínios

Como dito anteriormente, os domínios determinam características específicas de cada proteína e uma proteína pode conter um ou mais domínios. Logo, a comparação da estrutura de domínios da seqüência entrada pelo usuário e dos domínios dos "best hits" encontrados pelo BLAST dão ao usuário informação adicional à pesquisa BLAST inicial. De modo que ele possa avaliar se o “hit” encontrado na busca inicial realmente

representa uma seqüência homóloga à seqüência "query" dada, permitindo um aprimoramento do resultado final da anotação.

5.2.1. Implementação

Esta funcionalidade utiliza o aplicativo RPS-BLAST (Altschul et al, 1997) para fazer a procura por domínios da base CDD (Marchler-Bauer et al, 2005) na seqüência "query" e nas seqüências encontradas como "best hits" na pesquisa BLAST inicialmente realizada.

Entretanto, para as bases secundárias utilizadas na ferramenta já se tem o resultado precomputado dessa busca. O resultado da pesquisa RPS-BLAST das seqüências das bases contra a base de domínios CDD foi processado e incluído em tabelas do MySQL. Assim, quando o usuário ativa a funcionalidade de análise de estrutura de domínios da PCT, ao invés de se realizar uma comparação utilizando o RPS-BLAST, busca-se na base MySQL os resultados para os "best hits" encontrados.

Uma exceção ocorre para a base GOA, devido ao seu tamanho e ao da base CDD, o resultado da pesquisa RPS-BLAST de uma contra a outra iria gerar uma quantidade de dados considerada excessivamente grande. Assim, para os "best hits" da base GOA, bem como para a própria seqüência teste entrada pelo usuário, o RPS-BLAST é realizado "on the fly", fazendo com que o tempo de processamento da PCT para essa funcionalidade aumente um pouco.

No arquivo de configuração da ferramenta, caso a base secundária possua também uma base de domínios precomputados, isso deve ser especificado na estrutura de variáveis de configuração. No campo "bio_db[x][10]" (onde "x" é o número da base dada) se o valor for vazio (""), a ferramenta assume que não há uma tabela com os valores precomputados e faz a pesquisa RPS-BLAST na hora. Caso contrário, a ferramenta irá buscar na tabela os domínios para a seqüência encontrada como "best

hit", usando nessa busca o identificador da seqüência retornado pelo resultado do BLAST.

No caso de a base não possuir uma tabela com os valores precomputados, a ferramenta realizará então a pesquisa RPS-BLAST. O resultado dessa busca é armazenado num arquivo no disco, juntamente com outros arquivos intermediários gerados ao longo da execução da ferramenta. O formato do nome do arquivo é “XXX.bh.goa.rps” onde o “XXX” representa o número identificador da execução da ferramenta. Desse arquivo então são retiradas as informações dos domínios encontrados e o resultado é exibido conforme o modo de execução da ferramenta.

5.3. Análise filogenética

Com os resultados encontrados na pesquisa BLAST inicial, essa funcionalidade permite a criação de uma árvore filogenética para análise comparativa dos resultados.

Ela faz uso dos programas do pacote Phylip[21] (Felsenstein, J., 2005) que inferem a filogenia entre as seqüências e então cria uma arvore mostrando as relações filogenéticas encontradas.

Esses programas geram uma matriz de distância para a seqüência teste e todos os “hits” encontrados pelo BLAST, e a partir dessa matriz é gerada uma árvore filogenética.

Por meio dessa árvore, o usuário pode melhor avaliar se o “best hit” apontado pelo BLAST é realmente a seqüência que melhor identifica sua seqüência teste. Um exemplo de árvore desse tipo é exibido na Figura 11.

[21] Phylip - http://evolution.genetics.washington.edu/phylip.html

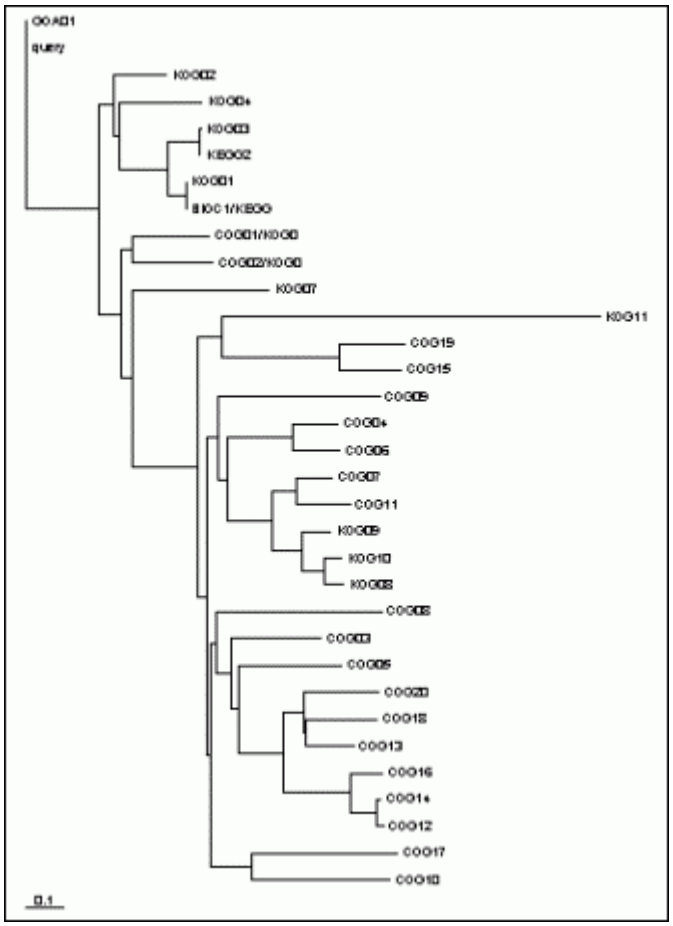


Figura 11 – Ilustração de uma árvore filogenética

5.3.1. Implementação

A funcionalidade de análise filogenética que gera uma árvore evolutiva entre os hits encontrados pela pesquisa BLAST contra as bases secundárias utiliza o programa “clustalw” e também os programas do pacote Phylip.

O primeiro passo é fazer um alinhamento global entre as seqüências onde foi encontrada similaridade com a seqüência “query” pelo BLAST e a própria seqüência “query”. Para isso as seqüências devem ser recuperadas de suas respectivas bases e reunidas em um único arquivo ao qual deve ser acrescentada a seqüência “query”.

Usando o comando “fastacmd” do pacote BLAST, as seqüências em que houveram “hits” são então retiradas das bases formatadas e a estas é acrescentada a seqüência de entrada. Feito isso, é necessário ainda que as linhas de identificação das seqüências sejam reescritas, pois elas não devem ter mais do que 10 caracteres, uma vez que o “clustalw”, software usado para gerar o alinhamento global exibe apenas 10 caracteres para identificar cada seqüência. Assim sendo, o identificador de cada seqüência é substituído por um identificador único, composto pelo nome da base de origem e um número que é atribuído seqüencialmente. Neste processo é preciso tomar cuidado para incluir somente uma cópia de cada seqüência da base “subject”, pois como o BLAST busca alinhamentos locais, para cada uma dessas seqüências pode ocorrer mais de um “hit” em trechos distintos. Esse identificador também é incluído na tabela onde são listados os resultados do BLAST, de modo que seja possível ao usuário comparar os resultados da análise filogenética com os resultados prévios da pesquisa BLAST.

Uma vez que o arquivo que contém as seqüências está no formato devido, usa-se o software “clustalw” para realizar o alinhamento múltiplo global das seqüências. Como resultado, obtém-se um arquivo contendo o alinhamento que será então usado como entrada para o “protdist”[22], programa do pacote Phylip que calcula uma matriz de distâncias para seqüências de aminoácidos.

A matriz de distâncias gerada pelo “protdist”, é então usada como entrada do “neighbor”, também do Phylip, que cria a árvore filogenética.

Por fim, são exibidos para o usuário os arquivos gerados pelo “neighbor”, contendo a árvore de filogenética em dois formatos: de forma gráfica e numa estrutura de parênteses.

[22] A versão do programa “protdist” usada nesse trabalho foi alterada por Sérgio Campos de modo a funcionar automaticamente, via parâmetros na linha de comando. A versão original era interativa e pedia os parâmetros ao longo da execução.

6. Modos de Uso da Ferramenta

A PCT apresenta dois modos de execução, um interativo onde o usuário primeiro submete a seqüência para a execução do BLAST contra as bases secundárias desejadas e então daí ele pode acionar outras opções de processamento. Já no modo de sumário o usuário já especifica todas as opções de execução logo no início e então todo processamento é feito de uma vez e em seguida é exibido um sumário dos resultados.

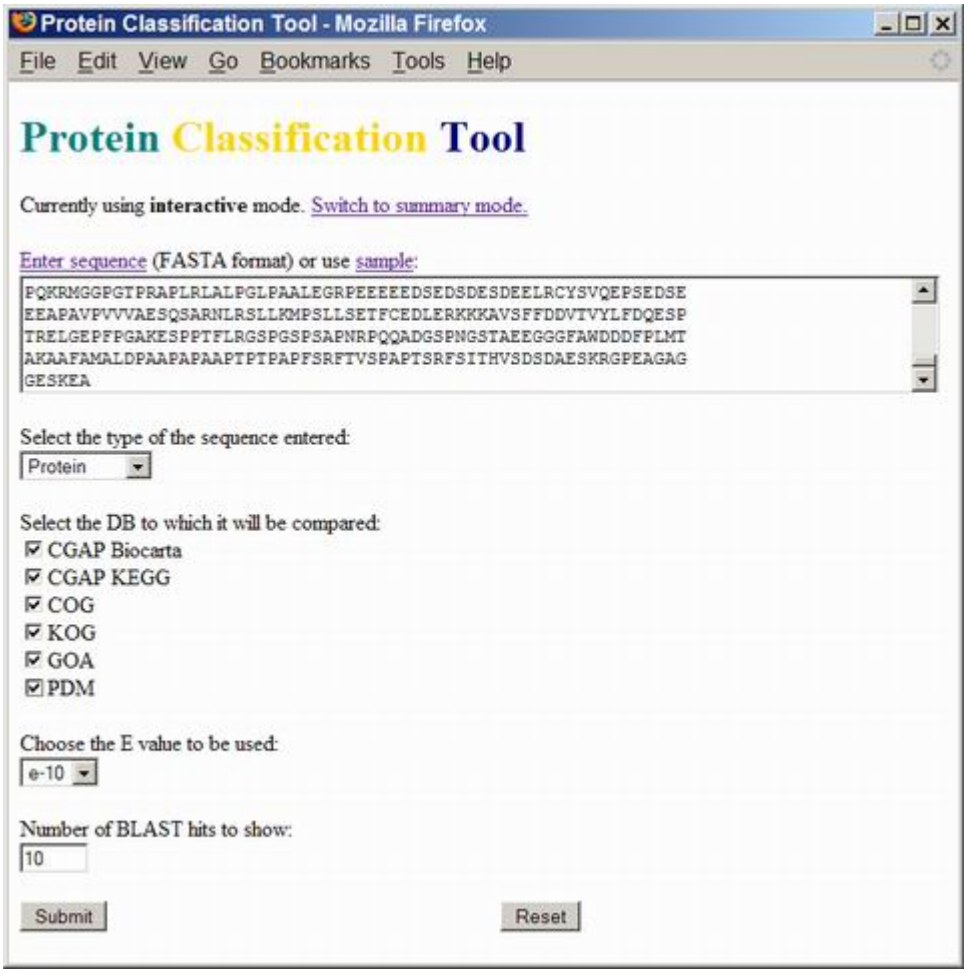


Figura 12 – Tela inicial da PCT

6.1. Modo interativo

No modo interativo, o primeiro passo é o usuário entrar com a seqüência a ser identificada, escolher a opção do tipo de seqüência correspondente, selecionar as bases secundárias contra quais o BLAST será executado e por último o valor de “E” (que corresponde a uma medida de confiabilidade do resultado da busca do BLAST). E então o usuário submete a busca à PCT. A Figura 12 mostra a página inicial da PCT.

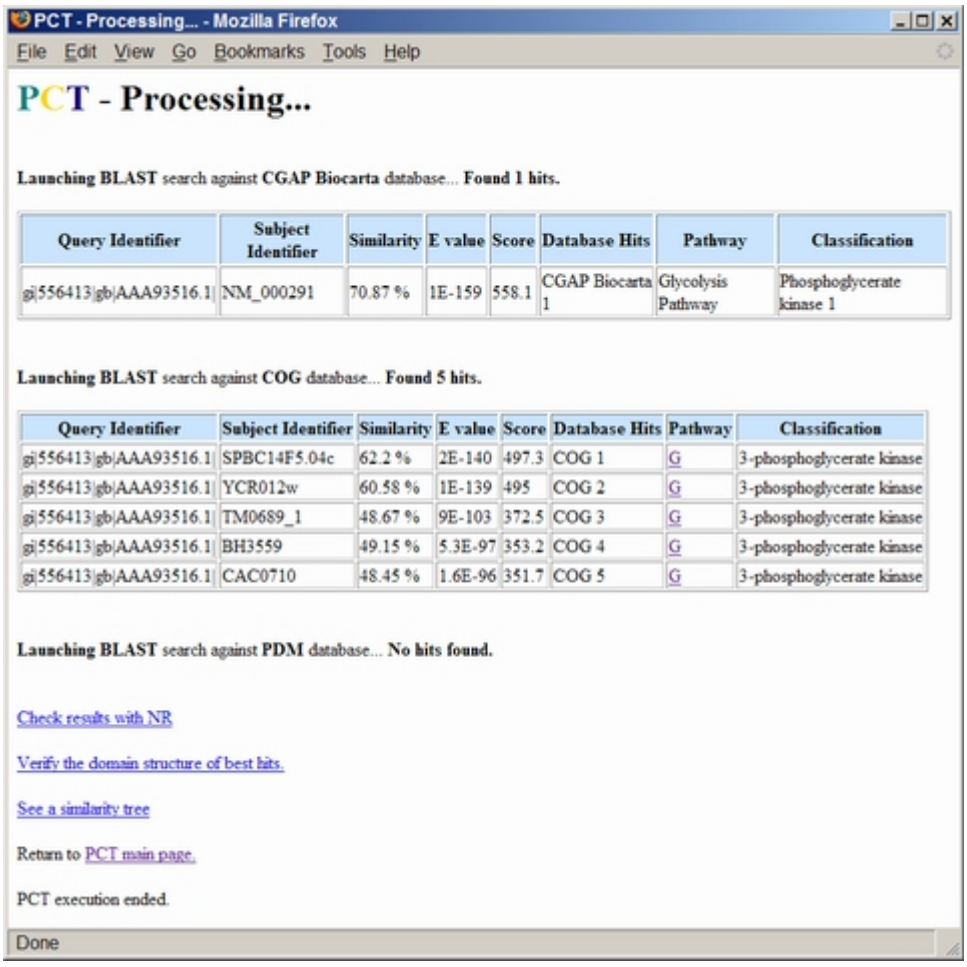
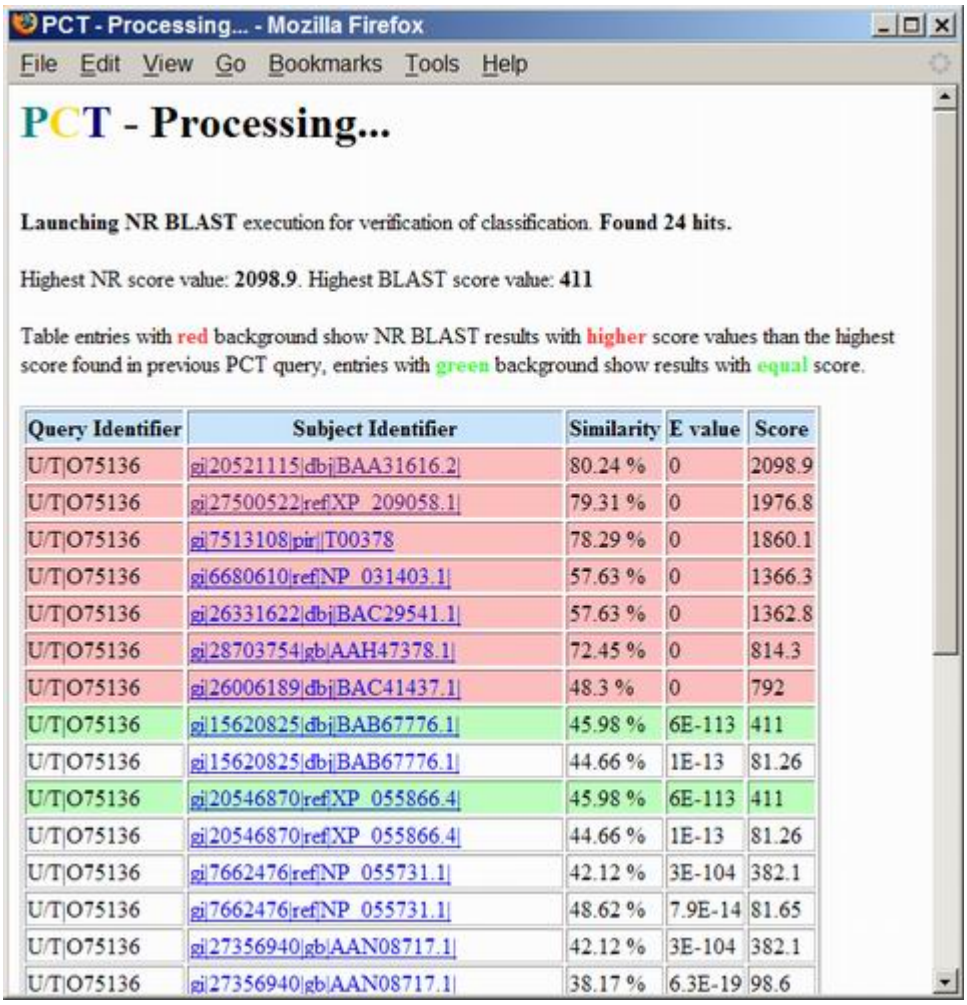


Figura 13 – Tela de resultados do modo interativo

Com esses dados, a PCT roda, internamente, um BLAST contra cada uma das bases selecionadas pelo usuário, levando em conta as opções por ele escolhidas.

Os resultados são então exibidos separados por base (Figura 13) e em seguida, caso haja “hits” do BLAST, o usuário tem três opções:

1) Rodar uma outra busca BLAST, mas dessa vez da seqüência de entrada contra a base NR.
2) Verificar a estrutura de domínios da seqüência de entrada
3) Gerar uma árvore de similaridade da seqüência de entrada juntamente com os “hits” encontrados



Query Identifier	Subject Identifier	Similarity	E value	Score
U/T|O75136	gi|20521115|dbj|BAA31616.2|	80.24 %	0	2098.9
U/T|O75136	gi|27500522|ref|XP_209058.1|	79.31 %	0	1976.8
U/T|O75136	gi|7513108|pir||T00378	78.29 %	0	1860.1
U/T|O75136	gi|6680610|ref|NP_031403.1|	57.63 %	0	1366.3
U/T|O75136	gi|26331622|dbj|BAC29541.1|	57.63 %	0	1362.8
U/T|O75136	gi|28703754|gb|AAH47378.1|	72.45 %	0	814.3
U/T|O75136	gi|26006189|dbj|BAC41437.1|	48.3 %	0	792
U/T|O75136	gi|15620825|dbj|BAB67776.1|	45.98 %	6E-113	411
U/T|O75136	gi|15620825|dbj|BAB67776.1|	44.66 %	1E-13	81.26
U/T|O75136	gi|20546870|ref|XP_055866.4|	45.98 %	6E-113	411
U/T|O75136	gi|20546870|ref|XP_055866.4|	44.66 %	1E-13	81.26
U/T|O75136	gi|7662476|ref|NP_055731.1|	42.12 %	3E-104	382.1
U/T|O75136	gi|7662476|ref|NP_055731.1|	48.62 %	7.9E-14	81.65
U/T|O75136	gi|27356940|gb|AAN08717.1|	42.12 %	3E-104	382.1
U/T|O75136	gi|27356940|gb|AAN08717.1|	38.17 %	6.3E-19	98.6

Figura 14 – Tela de resultados da comparação da “query” contra o NR

Caso não sejam encontrados “hits”, o usuário tem a opção de rodar um BLAST contra a base NR ou voltar pra página principal e fazer uma nova busca BLAST.

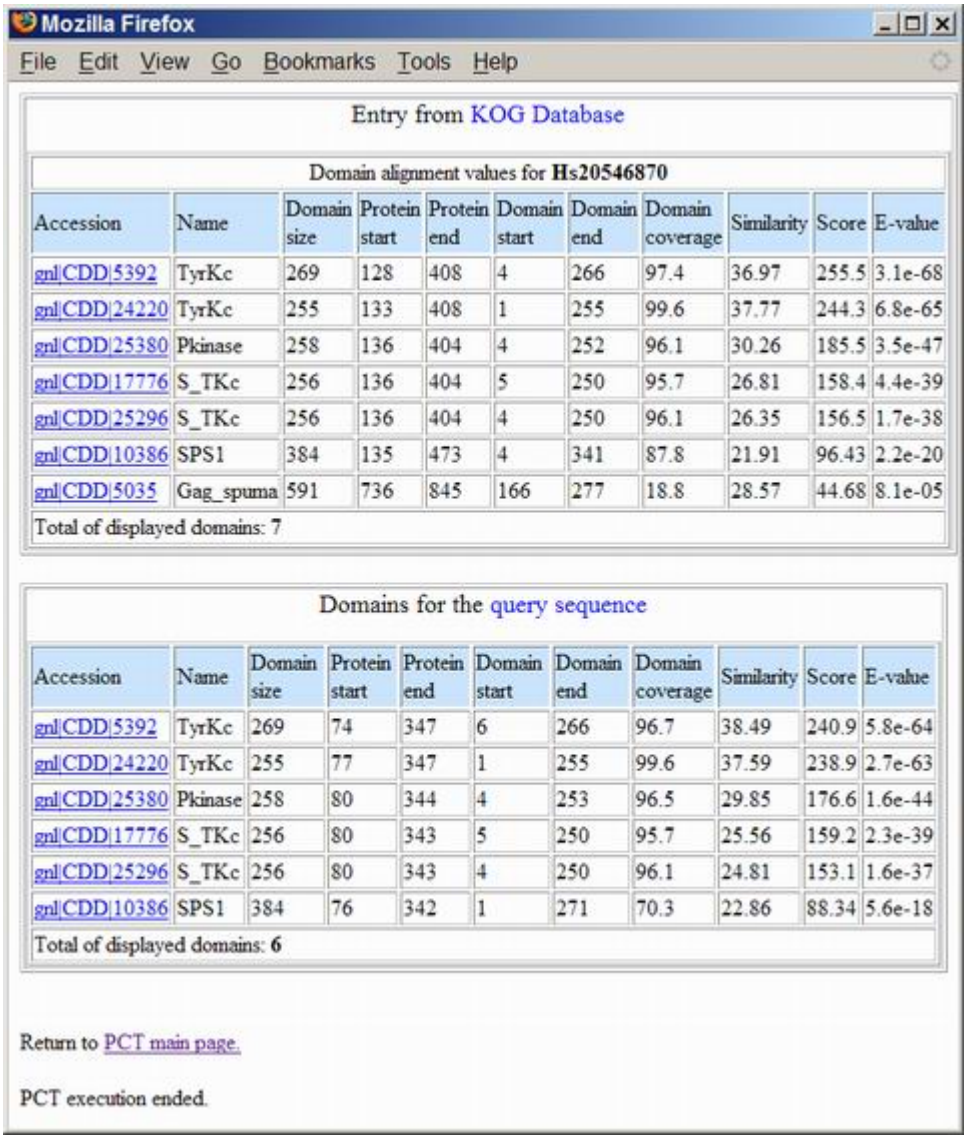


Figura 15 – Tela do resultado da funcionalidade de análise de dominios

Caso o usuário rode um BLAST contra a base NR, ele pode observar se os novos “hits” encontrados indicam que a anotação inicial foi imprecisa (Figura 14). Desse modo, se os novos resultados apresentarem valor de “score” igual ao valor máximo encontrado para as bases secundárias eles serão marcados em verde.

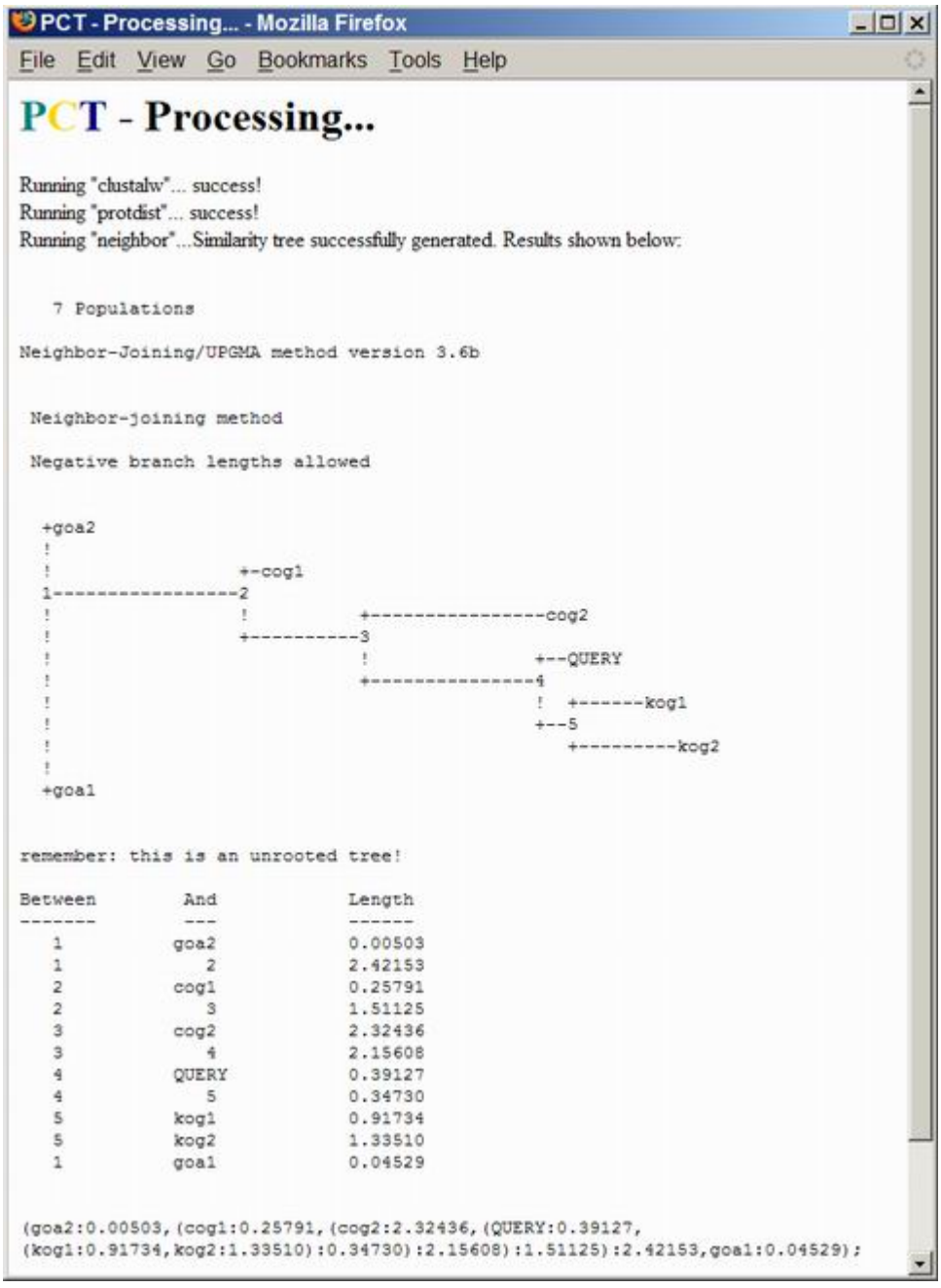


Figura 16 – Tela do resultado da funcionalidade de análise filogenética

Caso os novos valores de “score” excedam os valores previamente encontrados, estes serão marcados de vermelho, indicando que necessitam de atenção do usuário pois podem indicar que o resultado prévio não foi preciso, principalmente se a diferença entre o valor do “score” antigo e o novo forem muito discrepantes.

Outra opção do usuário após a busca BLAST inicial, é ver a estrutura de domínios da seqüência de entrada e dos “best hits” de cada base e por meio da comparação desses resultados, avaliar a qualidade da anotação (Figura 15).

Por fim, o usuário tem a opção de visualizar uma árvore de similaridade da seqüência de entrada juntamente com os “hits” das bases secundárias (Figura 16). Entretanto essa funcionalidade só está disponível para seqüências de proteínas, tanto em relação a seqüência de entrada quanto para as seqüências das bases. Assim, caso a seqüência de entrada seja de nucleotídeos, essa funcionalidade não estará disponível. As bases do CGAP também são de nucleotídeos e eventuais hits nessa base não entram na geração da árvore.

6.2. Modo de sumário

No modo de sumário o usuário entra com todas as opções logo de inicio e uma vez submetido ao processamento, a PCT já realiza todas as operações de uma só vez. O usuário verá então uma tela relatando o progresso da execução e uma vez concluído o processamento o usuário tem acesso a todos os resultados. A Figura 17 mostra a tela inicial do modo de sumário da PCT.

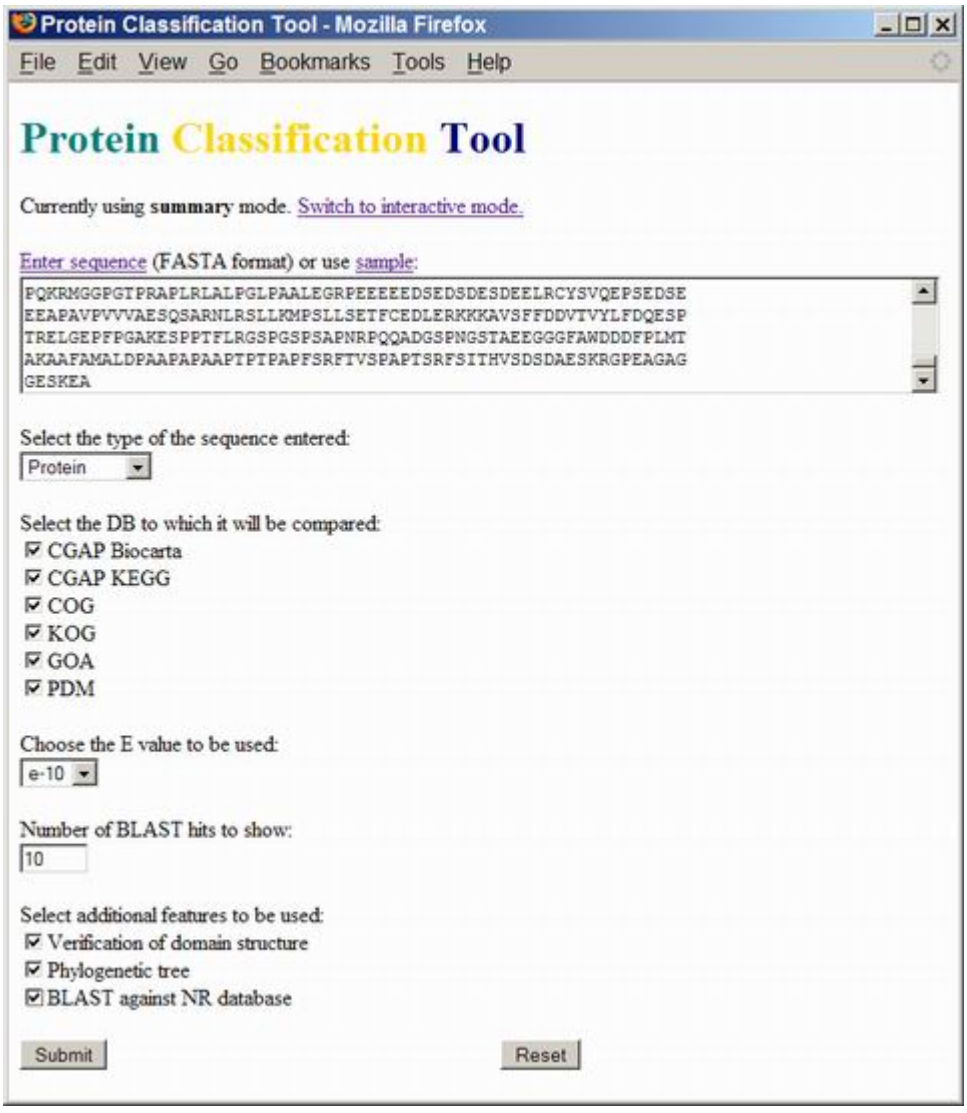


Figura 17 – Tela inicial do modo de sumário da PCT

Uma vez que o usuário submete o formulário inicial, a PCT irá realizar o processamento segundo as instruções do usuário. Será exibida uma nova página, onde será informado ao usuário o andamento da execução a medida que cada passo é executado e se ele termina com êxito, como pode ser visto na Figura 18.

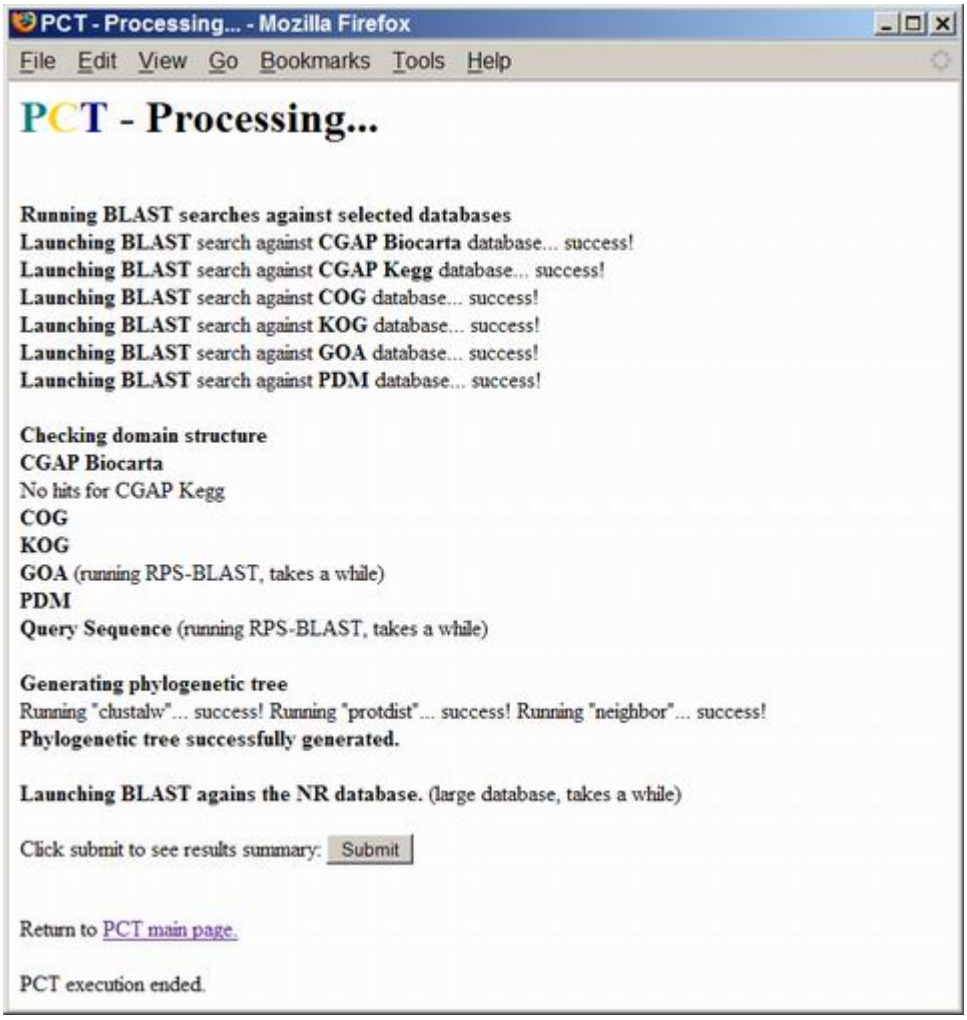
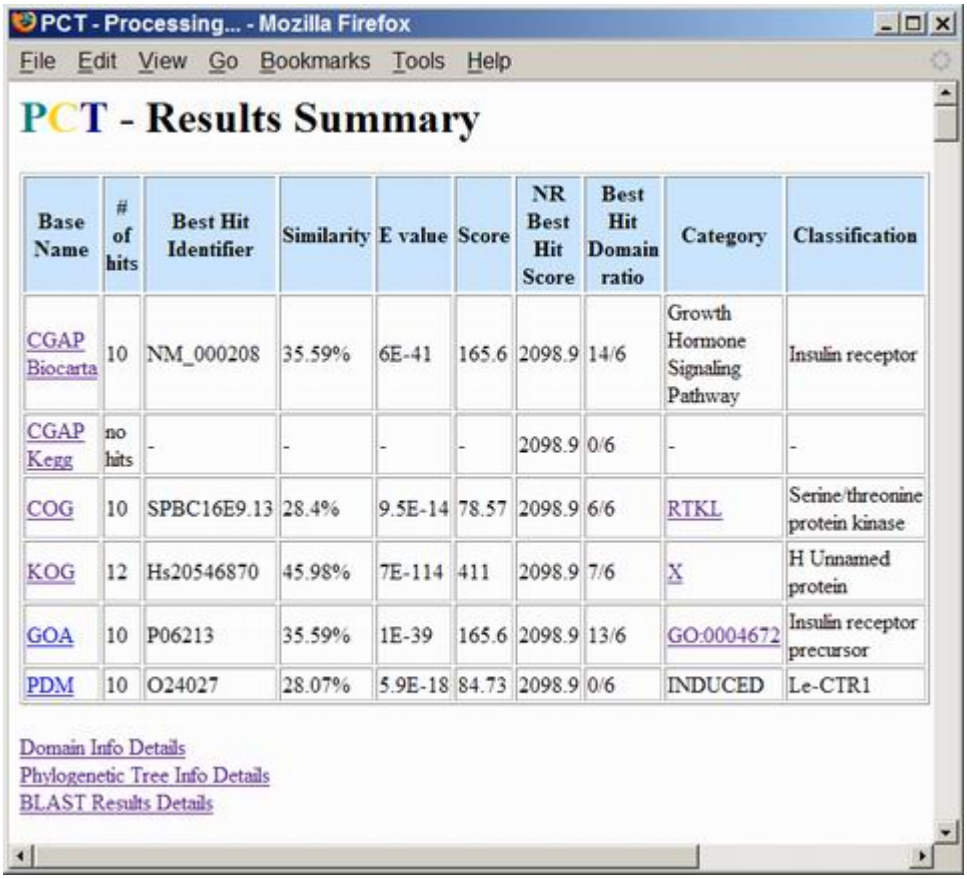


Figura 18 – Tela de processamento do modo de sumário

Ao fim do processamento, o usuário deve clicar no botão “Submit” que é exibido ao final das mensagens de execução, para então visualizar os resultados.

Os resultados são então exibidos todos na mesma página, sendo que primeiro há uma tabela que sumariza os resultados dos diversos métodos utilizados, como mostrado na Figura 19.



Base Name	# of hits	Best Hit Identifier	Similarity	E value	Score	NR Best Hit Score	Best Hit Domain ratio	Category	Classification
CGAP Biocarta	10	NM_000208	35.59%	6E-41	165.6	2098.9	14/6	Growth Hormone Signaling Pathway	Insulin receptor
CGAP Kegg	no hits	-	-	-	-	2098.9	0/6	-	-
COG	10	SPBC16E9.13	28.4%	9.5E-14	78.57	2098.9	6/6	RTKL	Serine/threonine protein kinase
KOG	12	Hs20546870	45.98%	7E-114	411	2098.9	7/6	X	H Unnamed protein
GOA	10	P06213	35.59%	1E-39	165.6	2098.9	13/6	GO:0004672	Insulin receptor precursor
PDM	10	O24027	28.07%	5.9E-18	84.73	2098.9	0/6	INDUCED	Le-CTR1

Figura 19 – Tela da PCT mostrando a tabela que sumariza os resultados

A primeira coluna exibe o nome da base secundária, sendo que no nome há um “link” para a parte inferior da página onde estão listados de forma completa os resultados do BLAST de cada base. Já a segunda coluna mostra o número de hits encontrados pelo BLAST, da terceira à sexta, respectivamente, são exibidos o identificador, taxa de similaridade, valor de “E” e “score” do alinhamento para o “best hit” de cada base. A sétima coluna exibe o mesmo valor em todas as linhas, ela mostra o “score” do “best hit” encontrado na comparação BLAST da seqüência “query” contra a base NR, a título de comparação com o “score” do “best hit” de cada uma das bases. A oitava coluna mostra o número de domínios encontrados no “best hit” da base

secundária em questão, em comparação com o número de domínios encontrado na seqüência “query”. Por fim, as duas últimas colunas mostram, respectivamente, a categoria dentro da base em que se encaixa o “best hit” e a classificação propriamente dita do “best hit” da base. Porém, tanto a coluna que mostra o “score” do “best hit” no NR quanto a que mostra a informação relativa aos domínios só serão exibidas quando a funcionalidade relativa a cada uma delas foi acionada no formulário que gerou a execução em questão.

Ao final dessa tabela, existem três “links” que apontam para posições inferiores nessa página de resultados, onde estão os resultados completos.

7. Estrutura para extensão da PCT

7.1. Incluindo bases de dados

Uma das vantagens da PCT é o fato do próprio usuário poder incluir novas bases de dados de seqüências a serem utilizadas nas pesquisas BLAST realizadas pela ferramenta.

Para cada base, tem-se um conjunto de seqüências associado a uma tabela MySQL (no formato mostrado acima). Como o acesso a acessas tabelas tem um núcleo comum padronizado, o usuário pode facilmente incluir outras bases para utilização na ferramenta.

Uma vez que o usuário tem as seqüências associadas a uma tabela MySQL, a inclusão na ferramenta se dá pela edição de 2 arquivos que compõem a estrutura da ferramenta (o arquivo “config.php”, que contém a informação relativa a cada base e o “index.php”, página inicial da ferramenta que referencia todas as bases). O processo de edição desses arquivos é absolutamente trivial, e é sugerido como trabalho futuro, a criação de uma interface para automatizar esse processo.

Devido à estrutura modular em que a ferramenta foi implementada, apenas com essas duas alterações a inclusão das bases refletirá em todas as funcionalidades da ferramenta.

No arquivo “config.php”, ficam as definições utilizadas pela ferramenta que o usuário pode alterar para adequá-la a seu sistema, como o diretório de execução da ferramenta, os caminhos dos executáveis utilizados e as informações relativas as bases de dados.

Essas definições são feitas via variáveis PHP e então o arquivo que as contém é incluído nos arquivos onde as variáveis são utilizadas. Abaixo um exemplo de arquivo “config.php”:

```
$bio_db_num = 2;                          // total number of bases

$bio_db[0][0] = "CGAP Biocarta";          // db display name
$bio_db[0][1] = "biocarta";               // db short name
$bio_db[0][2] = "db/biocarta_sequences.fasta";   // BLAST db
$bio_db[0][3] = "nucleotide";             // BLAST db type
$bio_db[0][4] = "biocarta";               // SQL table
$bio_db[0][5] = "Pathway";                // category column
$bio_db[0][6] = "";                       // category url
$bio_db[0][7] = "Classification";      // protein classification column
$bio_db[0][8] = "";                       // protein classification url
$bio_db[0][9] = "biocarta_cdd_tam";       // domain size SQL table
$bio_db[0][10] = "biocarta_cdd";          // domain hits SQL table

$bio_db[1][0] = "CGAP Kegg";              // db display name
$bio_db[1][1] = "kegg";                   // db short name
$bio_db[1][2] = "db/kegg_prot.fasta";     // BLAST db
$bio_db[1][3] = "nucleotide";             // BLAST db type
$bio_db[1][4] = "kegg";                   // SQL table
$bio_db[1][5] = "Pathway";                // category column
$bio_db[1][6] = "";                       // category url
$bio_db[1][7] = "Classification";      // protein classification column
$bio_db[1][8] = "";                       // protein classification url
$bio_db[1][9] = "kegg_cdd_tam";           // domain size SQL table
$bio_db[1][10] = "kegg_cdd";              // domain hits SQL table
```

No que se refere às bases de dados, define-se a variável “`$bio_db_num`” que contém o número de bases atualmente utilizadas.

Na matriz “`$bio_db[][]`” detalham-se as informações relativas a cada base. No trecho acima, dois exemplos de descrição de bases. O primeiro índice da matriz

indica o número identificador da base e o segundo índice determina uma informação relativa à base, como se segue:

0 – nome da base usado para exibição, pode conter mais de uma palavra

1 – nome da base em apenas uma palavra usado como identificador da base

2 – caminho a partir do diretório da ferramenta contendo o nome da base de seqüências formatadas para serem usadas no BLAST

3 – tipo de seqüências da base, pode ser “nucleotide” ou “protein”

4 – nome da tabela no MySQL

5 – nome da categoria em que a base classifica as proteínas a ser exibido na tabela de resultados

6 – URL a ser adicionada na tabela de resultados da base na coluna da categoria

7 – nome a ser usado na coluna em que é exibida a identificação da proteína

8 – URL a ser adicionada na tabela de resultados do BLAST para a base para a coluna da identificação da proteína

9 – nome da tabela MySQL onde ficam as informações relativas aos tamanhos dos domínios encontrados para cada seqüência da base

10 – nome da tabela MySQL onde ficam as informações relativas aos domínios encontrados para cada seqüência da base

7.2. Inclusão de novas funcionalidades

Visto que o problema da anotação demanda a associação de diversas técnicas para que se obtenha um bom resultado, e que novas técnicas estão sempre surgindo, a PCT foi concebida de modo a possuir um “design” modular. Permitindo, desta forma, que sua estrutura seja expandida com o mínimo de trabalho possível e permitindo a integração completa das novas funcionalidades com a estrutura existente.

As definições das bases secundárias utilizadas, ficam definidas no arquivo “config.php”, já as funções utilizadas em todo o processamento ficam no módulo “pct.php”. Assim,

No modo interativo da ferramenta, o primeiro passo no processo de execução é o BLAST contra as bases secundárias e a partir daí o usuário pode escolher se ele quer ou não utilizar alguma outra funcionalidade. Desta maneira a execução do BLAST e a análise dos resultados é feita pelo módulo “proc.php”, que então faz chamada para as outras funcionalidades, quando for o caso: “tree.php” para a funcionalidade de análise filogenética e “domains.php” para a análise de domínios.

Já no modo de sumário, todas as opções da ferramenta são exibidas logo na página de entrada e o usuário faz todas as suas escolhas antes de iniciar a execução.

Assim, a adição de novas funcionalidades deve ser feita de modo a integrar o novo processo à estrutura existente da ferramenta. Ainda que tenha sido feito o possível para facilitar o processo de integração de uma nova funcionalidade na estrutura pré-existente, é necessário o conhecimento de PHP e HTML de modo que o usuário possa tanto programar a nova funcionalidade como integrá-la na ferramenta.

A PCT deixa todos os arquivos intermediários criados no processamento no diretório “/tmp_files” dentro do diretório onde a PCT foi instalada. Assim, uma nova funcionalidade poderia utilizar os dados computados pela PCT.

Toda execução da PCT utiliza um identificador único que é dado pela hora exata da submissão do formulário inicial, esse identificador é obtido com o uso de uma função do PHP que retorna o tempo em segundos decorrido desde a “UNIX epoch” (1 de janeiro de 1970, 00:00:00 GMT). Assim, esse valor é usado como prefixo para os arquivos que conterão os dados de cada uma das execuções da PCT.

O primeiro arquivo armazenado é o que contém a seqüência entrada pelo usuário, ele é gravado com o sufixo “.txt”, assim para uma execução que se deu no momento “1155311456” (esse identificador será usado pra exemplificar os outros casos), o arquivo contendo a seqüência entrada pelo usuário teria o nome de “1155311456.txt”.

Em seqüência, o resultado de cada pesquisa BLAST realizada contra as bases secundárias é também armazenado, e para cada base o sufixo usado é “.out.base”, onde base é substituído pelo nome definido no arquivo “config.php” na variável de índice “1” da estrutura que armazena as informações de cada base. Assim, para a saída do BLAST contra a base “CGAP BioCarta” o nome do arquivo seria “1155311456.out.biocarta”.

No BLAST contra a base NR, o resultado é armazenado no arquivo “1155311456.txt.nr”

Para a funcionalidade de verificação da estrutura de domínios, os dados da maioria das bases secundárias está no banco de dados MySQL, mas para a seqüência de entrada (query) e para a base GOA, é executado na hora um RPS-BLAST contra a base de domínios CDD e o resultado é armazenado num arquivo contendo “.rps.base” como sufixo. Para a seqüência de entrada o sufixo é “.rps.query”, resultando em “1155311456.rps.query”, no caso da execução que estamos usando de exemplo.

Por fim, para a funcionalidade de análise filogenética, os primeiros arquivos gerados são os que contêm os identificadores dos “hits” encontrados para cada uma das bases secundárias. Eles são armazenados em arquivos cujo nome tem o sufixo “.ids.base”, novamente “base” substitui o nome da base secundária em questão.

A partir dos identificadores gravados nesse arquivo, as seqüências relativas a eles são então recuperadas das respectivas bases formatadas usadas no BLAST. Utilizando o programa “fastacmd”, do pacote BLAST, as seqüências são obtidas e então gravadas num arquivo com o sufixo “.fasta.base”.

A partir dos arquivos que contêm as seqüências para cada base, é gerado um novo arquivo que contém todas as seqüências, mas cada seqüência é reescrita de modo que sua linha de identificação tenha somente 10 caracteres, pois o processamento seguinte limita a linha de identificação de cada seqüência a esse tamanho. Esse novo arquivo que contém todas as seqüências tem o sufixo “.fasta”.

Esse arquivo com o sufixo “.fasta” é então usado como entrada para o “clustalw”, programa que fará o alinhamento global das seqüências, o primeiro passo para se gerar a árvore filogenética.
O resultado do alinhamento global é então gravado num arquivo com o sufixo “.phy”. Esse arquivo é usado então como entrada para o programa “protdist” que gera a partir do alinhamento global das seqüências uma matriz de distâncias e coloca a saída no arquivo com o sufixo “.prot”.

Por fim o programa “neighbor” usa o arquivo “.prot” como entrada e gera a árvore filogenética. A saída do “neighbor” são dois arquivos, um com sufixo “.nei” e outro com o sufixo “.tre”. O primeiro contém o desenho da árvore e o segundo uma outra construção da árvore com uma estrutura de parênteses.

Sumarizando, durante uma execução da PCT onde são ativadas todas as funcionalidades, os arquivos que são gerados são os seguintes (para uma execução com o identificador “1155311456”):

1155311456.fasta
1155311456.fasta.base
1155311456.ids.base
1155311456.nei
1155311456.out.base
1155311456.prot
1155311456.phy
1155311456.tre
1155311456.txt

8. Resultados

Com o intuito de validar a PCT e sua funcionalidade de anotação, executou-se dois conjuntos de testes. O primeiro conjunto foi executado com o intuito de avaliar e validar o funcionamento das funcionalidades da PCT, enquanto o segundo conjunto foi realizado com a finalidade de comparar os resultados da anotação obtida pela PCT com o de outras ferramentas.

8.1. Primeiro conjunto de testes

O primeiro conjunto de testes foi realizado com o intuito de validar o funcionamento da PCT e suas funcionalidades de anotação. Foram usadas como entrada 50 seqüências. Das 50 seqüências usadas, 14 são de bactéria (Escherichia coli DH10B) e 36 humanas (Homo sapiens), sendo ambas da via da “Glicólise” e considera-se que estão corretamente anotadas.

Para cada uma das seqüências foi realizada uma execução da PCT usando o modo de sumário. Essa execução consistiu em uma comparação BLAST da seqüência de entrada contra uma base secundária (KOG para as seqüências de bactéria e COG para as seqüências humanas), seguida pela execução das funcionalidades de verificação da estrutura de domínios, geração de árvore de similaridade e por fim uma comparação BLAST da seqüência de entrada contra a base NR. Nas Tabelas 5 e 6 são sumarizados os resultados encontrados para as 14 seqüências de bactéria.

Das 14 seqüências de bactéria testadas, 12 apresentaram “hits” no BLAST contra a base KOG, sendo o “best hit” capaz de prover a identificação e categorização funcional da proteína em 85,71% do total de casos (12 em 14) e 100% dos 12 casos onde houve “hits”. É comum encontrar parálogos na via glicolítica e nesse grupo de testes encontramos 2 cópias da “fructose-bisphosphate aldolase”, 3 da “phosphoglyceromutase” e 2 da “pyruvate kinase”. Entretanto as diferenças entre eles

não puderam ser identificadas nesse teste, o que já seria esperado, dada a distância evolutiva entre os organismos.

Seqüência	Query ID	Best Hit ID
1	glucokinase	No hit
2	glucosephosphate isomerase	Glucose-6-phosphate isomerase
3	6-phosphofructokinase I	Pyrophosphate-dependent phosphofructo-1-kinase
4	fructose-bisphosphate aldolase, class II	Fructose 1,6-bisphosphate aldolase
5	fructose-bisphosphate aldolase class I	No hit
6	triosephosphate isomerase	Triosephosphate isomerase
7	glyceraldehyde-3-phosphate dehydrogenase A	Glyceraldehyde 3-phosphate dehydrogenase
8	phosphoglycerate kinase	3-phosphoglycerate kinase
9	phosphoglyceromutase 1	Phosphoglycerate mutase
10	phosphoglyceromutase 2, co-factor independent	Phosphoglycerate mutase
11	phosphoglycero mutase III, cofactor-independent	Phosphoglycerate mutase
12	enolase	Enolase
13	pyruvate kinase I	Pyruvate kinase
14	pyruvate kinase II	Pyruvate kinase

Tabela 5 – Resultados: identificação funcional de seqüências de bactéria

A verificação dos domínios permitiu obter uma confirmação positiva da anotação em 75% dos casos (9 em 12) onde houve “hits” (na ausência de “hits”, não é possível comparar domínios. Na coluna “Domínios” da Tabela 6, a informação obtida pela comparação dos domínios é sumarizada na forma X/Y/Z, onde X é o número de domínios encontrados na seqüência de entrada, Y o número de domínios do “hit” analisado e Z o número de domínios coincidentes entre as duas seqüências. Nessa análise só foram aceitos domínios com similaridade acima de 45%, sendo que o número total de domínios encontrados para cada seqüência foi maior em alguns casos, mas estes foram descartados devido a baixa qualidade da identificação (Santos et al, 2008).

A pesquisa BLAST das seqüências contra a base NR permitiu confirmar o resultado encontrado para as 12 seqüências que apresentaram “hits” no BLAST contra

o KOG. Já para as 2 seqüências (1 e 5) que não haviam apresentado “hits” no BLAST contra o KOG, foi possível identificá-las corretamente.

Seqüência	Identidade	BLAST Score	Domínios	NR Ident	NR Score	NR Best Hit ID
1	-	-	2/0/0	95,64%	620,2	glucokinase
2	64,85%	722,2	2/2/2	100,00%	1116,3	glucose-6-phosphate isomerase
3	40,97%	205,3	3/3/3	95,00%	605,9	6-phosphofructokinase
4	50,14%	365,2	2/3/2	100,00%	728,4	fructose-bisphosphate aldolase
5	-	-	0/0/0	96,57%	673,3	fructose-bisphosphate aldolase
6	45,82%	200,3	2/2/2	100,00%	467,6	triosephosphate isomerase
7	70,34%	457,6	3/3/3	96,07%	631,7	glyceraldehyde-3-phosphate dehydrogenase
8	45,09%	308,5	1/1/1	100,00%	746,5	phosphoglycerate kinase
9	58,50%	306,6	2/2/2	100,00%	477,6	phosphoglyceromutase
10	35,27%	109	0/0/0	100,00%	430,3	phosphoglycerate mutase
11	45,40%	431,4	3/3/3	100,00%	1016,9	phosphoglycero mutase III, cofactor-independent
12	55,48%	433,3	3/3/3	93,06%	785,8	phosphopyruvate hydratase
13	48,68%	401,4	2/0/0	100,00%	911,0	pyruvate kinase
14	38,06%	271,6	3/0/0	96,88%	889,0	pyruvate kinase II

Tabela 6 - Resultados: estatísticas BLAST para seqüências de bactéria

As 36 seqüências humanas testadas apresentaram resultados semelhantes, onde a PCT encontrou “hits” e foi capaz de prover a identificação e classificação funcional corretas em todos os casos, porém, nesse caso também sem identificação de isoformas dos parálogos. As Tabelas 7 e 8 sumarizam os resultados.

Seqüências	Query ID	Best Hit ID
1	hexokinase 1 isoform HKI-td	Hexokinase
2	hexokinase 1 isoform HKI-ta/tb	Hexokinase
3	hexokinase 1 isoform HKI-ta/tb	Hexokinase
4	hexokinase 1 isoform HKI-R	Hexokinase
5	hexokinase 1 isoform HKI	Hexokinase
6	hexokinase 2	Hexokinase
7	hexokinase 3	Hexokinase
8	glucokinase isoform 1	Hexokinase
9	glucokinase isoform 2	Hexokinase
10	glucokinase isoform 3	Hexokinase
11	glucose phosphate isomerase	Glucose-6-phosphate isomerase
12	liver phosphofructokinase isoform a	6-phosphofructokinase
13	liver phosphofructokinase isoform b	6-phosphofructokinase
14	phosphofructokinase, muscle	6-phosphofructokinase
15	phosphofructokinase, platelet	6-phosphofructokinase
16	aldolase A	Fructose-1,6-bisphosphate aldolase
17	aldolase A	Fructose-1,6-bisphosphate aldolase
18	aldolase A	Fructose-1,6-bisphosphate aldolase
19	aldolase B	Fructose-1,6-bisphosphate aldolase
20	fructose-bisphosphate aldolase C	Fructose-1,6-bisphosphate aldolase
21	fructose-bisphosphate aldolase C	Fructose-1,6-bisphosphate aldolase
22	triosephosphate isomerase 1	Triosephosphate isomerase
23	glyceraldehyde-3-phosphate dehydrogenase	Glyceraldehyde-3-phosphate dehydrogenase/erythrose-4-phosphate dehydrogenase
24	phosphoglycerate kinase 1	3-phosphoglycerate kinase
25	phosphoglycerate kinase 2	3-phosphoglycerate kinase
26	2,3-bisphosphoglycerate mutase	Phosphoglycerate mutase 1
27	2,3-bisphosphoglycerate mutase	Phosphoglycerate mutase 1
28	enolase 1	Enolase
29	enolase 2	Enolase
30	enolase 3	Enolase
31	enolase 3	Enolase
32	pyruvate kinase, muscle isoform 1	Pyruvate kinase
33	pyruvate kinase, muscle isoform 2	Pyruvate kinase
34	pyruvate kinase, muscle isoform 2	Pyruvate kinase
35	pyruvate kinase, liver and RBC isoform 1	Pyruvate kinase
36	pyruvate kinase, liver and RBC isoform 2	Pyruvate kinase

Tabela 7 - Resultados: identificação funcional de seqüências humanas

Seq.	Ident.	BLAST Score	Dom.	NR Ident.	NR Score	NR Best Hit ID
1	36.16%	261.2	2/3/2	100,00%	1805.4	hexokinase 1 isoform HKI-td
2	36.16%	261.2	2/3/2	100,00%	1836.6	hexokinase 1 isoform HKI-ta/tb
3	36.16%	261.2	2/3/2	100,00%	1836.6	hexokinase 1 isoform HKI-ta/tb
4	36.16%	261.2	2/3/2	100,00%	1832.8	hexokinase 1 isoform HKI-R
5	35.94%	258.8	2/3/2	100,00%	1828.9	hexokinase 1 isoform HKI
6	37.36%	275.4	2/3/2	100,00%	1843.6	hexokinase 2
7	36.5%	273.5	2/3/2	95.12%	1725.3	hexokinase 3
8	34.63%	277.7	2/3/2	97.85%	911	glucokinase isoform 1
9	35.32%	276.6	2/3/2	97.85%	911.4	glucokinase isoform 2
10	35.32%	276.6	2/3/2	97.85%	912.1	glucokinase isoform 3
11	66.05%	727.6	2/2/2	100.00%	1141.3	glucose phosphate isomerase
12	45.61%	636	3/3/3	98.55%	1633.6	liver phosphofructokinase isoform b
13	46.56%	653.3	3/3/3	98.46%	1528.5	liver-type 1-phosphofructokinase
14	46.68%	651.7	3/3/3	98.21%	1539.6	phosphofructokinase, muscle
15	47.14%	657.5	3/3/3	100.00%	1567	phosphofructokinase, platelet
16	54.05%	327	4/4/4	95.05%	692.6	aldolase A
17	54.05%	327	4/4/4	95.05%	692.6	aldolase A
18	54.05%	327	4/4/4	95.05%	692.6	aldolase A
19	50.9%	311.2	4/4/4	96.7%	703	Fructose-bisphosphate aldolase B
20	52.52%	325.9	4/4/4	95,88%	696	fructose-bisphosphate aldolase C
21	52.52%	325.9	4/4/4	95.88%	696	fructose-bisphosphate aldolase C
22	53.25%	269.2	2/2/2	100.00%	502.7	triosephosphate isomerase 1
23	72.81%	493	3/3/3	100.00%	674.1	glyceraldehyde-3-phosphate dehydrogenase
24	66.83%	554.7	1/1/1	100.00%	827	phosphoglycerate kinase 1*
25	65.62%	541.6	1/1/1	100.00%	807	Phosphoglycerate kinase 2
26	51%	256.5	2/2/2	100.00%	531.2	2,3-bisphosphoglycerate mutase
27	51%	256.5	2/2/2	100.00%	531.2	2,3-bisphosphoglycerate mutase
28	63.51%	554.7	3/3/3	100.00%	861.3	enolase 1
29	62.59%	550.8	3/3/3	100.00%	862.4	enolase 2
30	63.74%	557.4	3/3/3	97.24%	837	Enolase 3 (beta, muscle)
31	63.74%	557.4	3/3/3	97.24%	837	Enolase 3 (beta, muscle)
32	54.12%	497.7	3/3/3	100.00%	1050.4	Pyruvate kinase isozymes M1/M2
33	52.82%	490	3/3/3	99.81%	1045	pyruvate kinase 3*
34	52.82%	490	3/3/3	99.81%	1045	pyruvate kinase 3*
35	52.83%	490.3	3/3/3	97.39%	1082.4	pyruvate kinase PK-R isoenzyme
36	52.83%	490.3	3/3/3	97.24%	1025	pyruvate kinase PK-L isoenzyme

Tabela 8 - Resultados: estatísticas BLAST para seqüências humanas

As seqüências 8, 9 e 10, originalmente anotadas como isoformas da “glucokinase” são identificadas como “hexokinase”, o que aparentemente seria incorreto, mas visto que a “hexokinase” é a isoforma 4 da “glucokinase”, tem-se um resultado correto, porém com identificação incorreta da isoforma.

Nas seqüências de 1 a 10, a coluna de domínios mostra o resultado “2/3/2”, que também é considerado positivo, visto que os dois domínios encontrados na seqüência de entrada são mapeados no “hit” do BLAST.

Enfim, para as seqüências humanas a PCT conseguiu sucesso na identificação e categorização funcional em 100% dos casos, mas em nenhum deles foi capaz de determinar a isoforma da proteína, o que demandaria análise mais minuciosa e possivelmente comparação com outras bases de seqüências.

Esse conjunto de testes nos permitiu observar que a PCT teve um desempenho muito bom no processo de anotação de seqüências. Ainda que tenha falhado em determinar as isoformas das proteínas em alguns casos, o resultado final foi satisfatório.

8.2. Segundo conjunto de testes

Um segundo conjunto de testes foi realizado com o intuito de comparar o desempenho da PCT com o de outras ferramentas no processo de anotação. Optou-se por usar o Blast2GO como ferramenta a ser comparada, dada a estabilidade do serviço e a disponibilidade de funcionalidades, que se mostrou a melhor entre as opções disponíveis (como discutido no Capítulo 4).

Nesse teste foram usadas 25 seqüências de nucleotídeos de Schistosoma mansoni que apresentaram algum “hit” contra a base KOG. O teste consistiu em comparar os resultados do processo de anotação usando a ferramenta Blast2GO com os resultados obtidos na PCT. Na PCT foram realizadas execuções no modo de sumário, usando o BLAST contra as bases KOG e GOA.

Analisamos os resultados obtidos pela PCT considerando parâmetros estabelecidos por Barbosa-Silva (Barbosa-Silva, 2008) para comparação de resultados na base KOG (ou a base GOA) com a base NR:

1) Score NR – Score KOG < 150
2) Score NR / Score KOG < 2
3) Identidade NR – Identidade KOG < 15

Cada um desses fatores seria um indicativo da qualidade do resultado encontrado no KOG, e sendo os três fatores favoráveis, pode se considerar um forte indício de que o resultado provido pelo KOG pode ser usado como uma anotação válida, a Tabela 9 sumariza esses resultados.

Seq.	Ident. KOG	Score KOG	Ident. GOA	Score GOA	Ident. NR	Score NR
1	85.53%	134.4	85.53%	132.9	86.76 %	100.5
2	42.42%	104.8	42.42%	104.8	42.42 %	104.8
3	43.79%	125.9	43.79%	125.9	43.79 %	125.9
4	97.67%	409.8	94.88%	397.1	93.01 %	410.6
5	71.79%	125.9	95.06%	157.1	95.06 %	157.1
6	80.56%	302	82.68%	306.2	82.78 %	308.1
7	30.6%	117.9	33.48%	121.3	98.25 %	455.3
8	29.74%	86.27	29.74%	85.89	30.74 %	87.81
9	55.1%	117.9	52.04%	108.6	57.14 %	105.5
10	53.61%	273.1	60.15%	304.3	84.58 %	421.8
11	37.97%	103.2	34.91%	103.2	44.44 %	104
12	81.01%	126.3	84.21%	131.7	83.56 %	110.2
13	40.32%	86.27	29.47%	70.86	94.74 %	357.8
14	42.48%	84.73	37.72%	79.72	42.48 %	84.73
15	35.59%	115.2	35.59%	115.2	35.59 %	115.2
16	52.66%	224.6	52.66%	226.1	53.62 %	234.6
17	66.19%	305.1	64.45%	304.3	65.9 %	313.5
18	56.06%	240.4	56.06%	240.4	56.06 %	240.4
19	56.35%	285.4	63.56%	328.2	91.16 %	401.7
20	40.85%	113.6	40.23%	109.4	47.06 %	131
21	31.03%	74.71	28%	82.8	32.48 %	84.73
22	63.12%	179.1	63.12%	181.4	68.6 %	178.3
23	38.33%	98.6	46.79%	94.36	48.33 %	117.5
24	33.54%	104.4	33.54%	104.4	33.54 %	95.9
25	53.76%	105.5	53.76%	105.5	53.76 %	105.5

Tabela 9 – Comparação de resultados entre as bases KOG, GOA e NR

Através dessa análise dos resultados obtidos pela base KOG, vemos que em 20 das 25 seqüências (80%) os 3 fatores são positivos. Em 1 (4%) 2 fatores são positivos, em 2 (8%), apenas 1 fator é positivo e em outras 2 (8%) nenhum dos fatores é positivo.

Comparando o resultado do KOG com o GOA, que é uma base maior e mais completa, o resultado melhora um pouco. Tendo 21 (84%) seqüências com 3 fatores positivos, 2 (8%) com 2 fatores e outras 2 (8%) onde nenhum dos fatores é positivo.

Quanto à classificação em categorias funcionais, a PCT e o Blast2GO apresentaram resultados bastante similares. Em 19 das seqüências (76%) o “GO term” apresentado pela PCT foi um dos termos apresentados pelo Blast2GO. Dos 6 restantes, em 1 deles (4%) o resultado apresentado pela PCT era um nodo “pai” de um termo apresentado pelo Blast2GO, e nos outros 5 (20%) os resultados eram nodos “primos” distantes na árvore do GO.

Por fim, a PCT apresentou “strings” de identificação melhores que o Blast2GO em boa parte das seqüências. Enquanto o Blast2GO obtém as suas strings de identificação a partir do NR, as strings da PCT vêm do KOG e do Uniprot, que provêem strings de identificação mais significativas. Em 12 seqüências (48%) a string provida pelo Blast2GO é pior que a provida pela PCT, sendo que em 3 casos (12%) não provê nenhuma informação útil (por exemplo, “1 homolog”) e em 9 casos (36%) provê menos informação que a string provida pela PCT. O restante dos casos as strings são muito parecidas ou se referem a diferentes seqüências.

Enfim, a partir dos resultados obtidos nesse conjunto de testes, podemos dizer que a anotação provida PCT apresentou qualidade comparável à provida pelo Blast2GO, sendo melhor em alguns aspectos, tal como a qualidade da “string” de identificação da seqüência.

Seq.	KOG	GOA	Blast2GO
1	Ubiquitin-like protein	Neddylin	neural precursor celldevelopmentally down-regulated 8
2	Glycosyltransferase	Alpha-1,3-mannosyltransferase ALG2	alg2 protein
3	Protein involved in plasmid maintenance/nuclear protein involved in lipid metabolism	Lipin 2	lipin 3
4	Ubiquitin and ubiquitin-like proteins	Ubiquitin	im:6892314 protein
5	Calmodulin and related proteins (EF-Hand superfamily)	Calmodulin beta (Fragment)	calmodulin
6	GTP-binding ADP-ribosylation factor Arf1	ADP-ribosylation factor 4	adp-ribosylation factor 4
7	Serpin	Squamous cell carcinoma antigen	squamous cell carcinoma antigen
8	Collagens (type IV and type XIII), and related proteins	Collagen alpha 2(I) chain precursor	fibrillar collagen
9	Nuclear transport receptor CRM1/MSN5 (importin beta superfamily)	Exportin 1	1 homolog
10	Myosin class II heavy chain	Myosin heavy chain, striated muscle	myosin heavy chain
11	ATP-dependent RNA helicase	ATP-dependent RNA helicase ded1	atp-dependent rna helicase of dead box family
12	Histone 2A	Histone H2A	histone h2a
13	Amidases	Fatty-acid amide hydrolase	elegans proteinpartially confirmed by transcript evidence
14	RNA polymerase III subunit C11	DNA-directed RNA polymerases III 12.5 kDa polypeptide	polymeraseiii (dna directed) polypeptide k
15	Deoxyribonuclease II	Deoxyribonuclease II precursor	deoxyribonuclease ii beta
16	Pyruvate carboxylase	Pyruvate carboxylase, mitochondrial precursor	pyruvate carboxylase
17	Glutamine synthetase	Glutamine synthetase	glutamine synthetase
18	Nucleoside diphosphate kinase	Nucleoside diphosphate kinase homolog 5	non-metastatic cellsprotein expressed in (nucleoside-diphosphate kinase)
19	Phosphoglycerate mutase	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase	phosphoglycerate mutase 1 family
20	Lipoyltransferase	Lipoyltransferase	lipoate-protein ligase
21	Prosaposin	Sulfated glycoprotein 1 precursor	mgc80725 protein
22	Manganese superoxide dismutase	Superoxide dismutase [Mn], mitochondrial precursor	superoxide dismutasemitochondrial
23	O-methyltransferase	Caffeoyl-CoA O-methyltransferase 2	o-family 3
24	P-type ATPase	Potential phospholipid-transporting ATPase IA	atpase ii
25	Protein DRE2, required for cell viability	Anamorsin	cytokine induced apoptosis inhibitor 1

Tabela 10 – Comparação das strings de classificação providas pela PCT e pelo Blast2GO

9. Trabalhos Futuros

Sugerem-se como trabalhos futuros a criação de um conjunto de scripts tanto para atualização automática das bases de dados utilizadas na PCT e quanto para automatizar a adição de bases customizadas. Uma implementação da ferramenta que fosse independente de um servidor web poderia facilitar a instalação e melhorar a performance.

O funcionamento da ferramenta também poderia ser alterado de modo a permitir o processamento de diversas seqüências ao mesmo tempo de maneira funcional, possivelmente armazenando os dados no banco de dados para análise posterior.

Os alinhamentos entre seqüências e domínios poderiam ser exibidos de maneira gráfica, de modo a auxiliar a visualização da cobertura do alinhamento.

A funcionalidade de árvores filogenéticas atualmente só está implementada para seqüências de aminoácidos. Isso porque os softwares do pacote Phylip que criam as matrizes de distâncias para alinhamentos globais não funcionam de maneira completamente automática, e é necessário alterar seu código para que ele possa ser utilizado na automação do processo. Isso foi feito para o “protdist”, que gera a matriz de distâncias para alinhamentos globais de proteínas, mas o mesmo deve ser feito para o “dnadist” que faz o mesmo para alinhamentos globais de seqüências de DNA, o que não foi feito nesse trabalho devido às restrições de prazo. Outra melhoria que pode ser adicionada é a associação da relação evolutiva entre espécies à árvore gerada a partir das seqüências, de modo a permitir uma analise filogenética mais completa.

Por fim, poderia-se buscar maneiras de automatizar ainda mais o processo de anotação, tentando combinar os resultados das diversas bases secundárias com os outros métodos (domínios e árvore filogenética) automaticamente e sugerindo ao usuário uma anotação baseada na comparação dessas informações.

10. Conclusão

Com esse trabalho pôde-se observar ganhos significativos no processo de anotação, à medida que a PCT provê diversas bases secundárias junto com outras funcionalidades de anotação. Isso faz com que a PCT seja capaz de fornecer um serviço de anotação com acurácia e praticidade, sem a necessidade de análises externas. Assim, vemos que a integração de diversas funcionalidades de anotação numa mesma ferramenta permite que o processo fique bem mais ágil e intuitivo, levando também a melhores resultados.

Ao longo do desenvolvimento desse trabalho, pudemos acompanhar o desenvolvimento de diferentes métodos de anotação e vimos que o surgimento de novas funcionalidades melhorou consideravelmente os resultados obtidos. Assim, uma grande virtude dessa ferramenta é permitir de maneira relativamente simples, que o próprio usuário altere seu funcionamento de modo a incluir novas funcionalidades, permitindo que a PCT continue atualizada, agregando novos métodos de anotação, ao contrário de ferramentas fechadas incapazes de incorporar novas bases ou funcionalidades.

11. Referências

Ainscough, R., Baroill, S. and Barlow, k. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science, 282: 2012-2018, 1998.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. Basic Local Alignment Search Tool. Journal of Molecular Biology, 215: 403-410, 1990.

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402, 1997.

Apweiler R., Bairoch A., Ferro S., Natale D. A., Yeh L.S. et al.. UniProt: the Universal Protein Knowledgebase. Nucleic Acids Res. 32: D115-D119, 2004.

Ashburner M., Ball, C. A., Blake., J. A. et al. Gene Ontology: Tool for the Unification of biology. The Gene Ontology Consortium. Nature Reviews Genetics, 25(1):25-9, 2000.

Barbosa-Silva, A., Mineração de texto, agrupamento de seqüências e integração de dados para o desenvolvimento da Plant Defense Mechanisms Database. Tese de Doutorado em Bioinformática-UFMG, 2008.

Bedell, J., Korf, Ian. e Yandell, M. BLAST. O'Reilly, 2003.

Bravo-Neto, E. Atualização e Teste de uma ferramenta de classificação de proteínas utilizando bases de dados secundárias. Monografia de conclusão de curso de Ciências Biológicas. ICB-UFMG, 2004.

Califano, A. Advances in Sequence Analysis. Current Opinion in Structural Biology, 11: 330-333, , 2001.

Celniker, S. E. The Drosophila Genome. Current Opinion in Genetics and Development, 10: 612-616, 2000.

Cho, Y. and Walbot, V. Computational methods for gene annotation: the Arabidopsis genome. Current Opinion in Biotechnology, 12(2):126-30, 2001.

Claverie, J.-M., Poirot, O. and Lopez, F. The Difficulty of Identifying Genes in Anonymous Vertebrate Sequences. Computers and Chemistry, 21: 203-214, 1997.

Collins, F., Guyer, M., Peterson, J., Fesseenfeld, A. and Wetterstrand, K. A. Initial Sequencing and Analysis of the Human Genome. Nature, 409: 860-921, 2001.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676, 2005.

Dávila, A. M., Lorenzini, D. M., Mendes, P. N., Satake, T. S., Sousa, G. R., Campos, L. M., Mazzoni, C. J., Wagner, G., Pires, P. F., Grisard, E. C., Cavalcanti, M. C., Campos, M. L., and Campos, M. L. GARSA: genomic analysis resources for sequence annotation. Bioinformatics 21, 23, 4302-4303., 2005

Doerks, T., Bairoch, A. and Bork, P. Protein Annotation: Detective work for Function Prediction. Trends in Genetics, 14: 248-250, 1998.

Eisenberg, D., Marcotte, E. M., Xenarios, I. and Yeates, T. O. Protein Function in the Post Genomic Era. Nature, 405: 823-826, 2000

Ewing, B., Hillier, L., Wendl, M.C. and Green, P. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res. 8: 175-185, 1998.

Ewing, B. and Green, P. Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res. 8: 186-194, 1998.

Faria-Campos, A. Bioinformática Aplicada a Caracterização de Genes: O Schistosoma mansoni como modelo. Tese de Doutorado do ICB-UFMG, 2005.

Felsenstein, J. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166, 1989.

Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, 2005.

Grayhack, E. J. and Phizicky, E. M. Genomic Analysis of Biochemical Function. Current Opinion in Chemical Biology, 5: 34-39, 2001.

Gibas, C. and Jambeck, P. Developing Bioinformatics Computer Skills, ed. LeJeune, L. O´Reilly, Sebastopol, 2001.

Harris M. A., Clark J., Ireland A., Lomax J., Ashburner M., Wood V., White R. et al; “Gene Ontology The Gene Ontology (GO) database and informatics resource”. Nucleic Acids Res. Jan 1;32 Database issue: D258-61, 2004.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M. The KEGG Resource for Deciphering the Genome. Nucleic Acids Research, 32(Database issue):D277-80, 2004.

Koski, L., Gray, M. W., B Franz Lang, F. B. and Burger, G. AutoFACT: An Annotation Functional and Classification Tool. BMC Bioinformatics, 6:151, 2005.

Kramer, R. and Cohen, D. Functional Genomics to New Drug Targets. Nature Reviews Drug Discovery, 3: 965-72, 2004.

Liu, E. T. Genomic Technologies and the Interrogation of the Transcriptome. Mechanisms of Ageing and Development, 126: 153-59, 2005.

Mathé C., Sagot M., Schiex T. and Rouzé P. Currento methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res., Vol. 30 No. 19 4103-4117, 2002

Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Bryant SH, et al. CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res., 33 Database Issue:D192-6, 2005

Miller, W., Makova, K. D., Nekrutenko, A. and Hardison, R. C. Comparative Genomic Schistosoma Annual Review of Genomics and Human Genetics, 5: 15-56, 2004.

Misra, S., Crosby, M. A., Mungall, C. J. and Matthews, B. B. Annotation of the Drosophila Melanogaster Euchromatic Genome: a Systematic Review. Genome Biology, 3, RESEARCH0083, 2002.

Natale, D. A., Shankaavaram, U. T., Galperin, M. Y., Wolf, Y. I. et al. Towards Understanding the First Genome Sequence of A Crenarcheon by Genome Annotation Using Clusters of Orthologous Groups of Proteins, COGs. Genome Biology, 1: 0009.1-0009.19, 2000.

Nemoto, Y. Trends of Genomic Researches for Drug Discoveries. Nippon Rinsho 56, 224-32, 1998.

Pollock, D. D. Genomic Biodiversity, Phylogenetics and Coevolution in Proteins. Applied Bioinformatics, 1: 81-92, 2002.

Santos, L.S., Prosdocimi, F., Ortega, J. M. Essential amino acid usage and evolutionary nutrigenomics of eukaryotes - insights for the differential usage of amino acids in protein domains and extra-domains. Genetics and Molecular Research, 2008.

Smith T.F. e Waterman M.S. Identification of common molecular subsequences. J. Mol. Biol. ;147:195–197, 1981.

Stein, L. Genome Annotation: from Sequence to Biology. Nature Reviews Genetics, 2: 493-503, 2001.

Tatusov, R. A., Fedorova, N. D., Jackson, J. D., Jacobs, A. R. et al. The COG Database: Un Updated Version Includes Eukaryotes. BMC Bioinformatics, 4: 41, 2003.

Venter, J. C., Adams, M. D., Myers, E. W. and Li, P. W. The Sequence of the Human Genome. Science, 291, 2001.

Yakunin, A. F., Yee, A. A., Savchenko, A., Edwards, A. M. and Arrowsmith, C. H. Structural Proteomics: A Tool for Genome Annotation. Current Opinion in Chemical Biology, 8: 42-48, 2004.

Whisstock, J.C. and Lesk, A.M. Prediction of Protein Function from Protein Sequence and Structure. Quartely Review of Biophysics, 36: 307-340, 2003.

12. Glossário

Aminoácidos – Aminoácidos são as moléculas que constituem as proteínas. Existem 20 diferentes tipo de aminoácidos e cada um deles é codificado no DNA por 3 bases segundo o código genético, sendo que para alguns aminoácidos existe mais de uma seqüência correspondente.

Base, base nitrogenada, par de bases – Bases nitrogenadas são compostos contendo nitrogênio que compõem as moléculas de DNA e RNA. No DNA são 4 os tipos diferentes de bases: adenina (A), citosina (C), guanina (G) e timina (T). No RNA a timina é substituída pela uracila (U). A expressão “par de bases” se refere ao fato de que no DNA a molécula tem uma estrutura de dupla hélice, e assim as bases de uma das fitas ficam ligadas às bases da outra por meio de ligações de hidrogênio e são então referidas como um par de bases.

Códon – Códon é o nome dado ao conjunto de 3 bases do DNA ou RNA. Um códon determina um aminoácido segundo o código genético. Sendo que se tem 4 bases diferentes na composição de uma molécula de DNA ou RNA, tem-se 64 combinações possíveis, entretanto existem apenas 20 aminoácidos o que implica que vários aminoácidos são determinados por diferentes combinações de bases.

Código genético – O código genético é a relação entre a seqüência de bases no DNA e a seqüência correspondente de aminoácidos que é empregada no processo de síntese de proteínas. Quase todos os seres vivos usam o mesmo código genético, chamado de código genético padrão, entretanto alguns poucos organismos utilizam pequenas variações desse padrão.

DNA – O ácido desoxirribonucléico (ADN, ou em inglês “deoxyribonucleic acid”, DNA), geralmente na forma de uma dupla hélice, é a molécula que contém as instruções genéticas que especificam o desenvolvimento biológico de todas as formas de vida e de vários vírus. O DNA é um longo polímero de nucleotídeos (ou um polinucleotídeo) e nele estão codificados por meio do código genético a seqüência de aminoácidos que determina todas as proteínas de um organismo.

Em células de eucariotos (tais como animais e plantas) a maior parte do DNA está no núcleo da célula, agrupado em uma ou mais macromoléculas chamadas cromossomos. Por outro lado, em organismos mais simples, os procariotos (tais como as bactérias), o DNA fica difuso no citoplasma da célula. Organelas celulares tais como cloroplastos e mitocôndrias também possuem DNA.

Domínios, domínios de proteínas ou domínios conservados – domínios podem ser vistos como unidades funcionais ou estruturais (freqüentemente ambas) de uma proteína, uma vez que a estrutura de uma proteína, em geral, determina sua função. Domínios costumam ser identificados como partes recorrentes da estrutura ou da seqüência de uma proteína que aparecem em contextos diversos.

Eucariotos – São organismos que possuem células eucarióticas, ou seja, células que possuem um núcleo verdadeiro, rodeado por uma membrana e também possuem outras organelas com membranas.

São organismos mais complexos que podem ser unicelulares ou multicelulares, o domínio "Eukaryota" é dividido em animais, plantas, fungos e protistas.

Nos eucariotos o DNA fica no núcleo da célula agrupado em um ou mais cromossomos e a freqüência de regiões codificantes tende a diminuir em proporção inversa a

complexidade do organismo sendo cerca de 70% na levedura, 25% na mosca e cerca de %5 no homem.

Filogenia – A filogenia trata das relações evolutivas entre diversos organismos, ou como referido nesse trabalho, entre seqüências homólogas.

Gene – Genes são as unidades da hereditariedade em seres vivos. Eles são codificados no material genético do organismo (em geral DNA ou RNA), e controlam o desenvolvimento e o comportamento do organismo. Durante a reprodução, o material genético é transmitido aos descendentes mas ele também pode ser transmitido entre organismos sem relação de parentesco, por exemplo, pelos vírus.

Genoma – O genoma de um organismo é o conjunto de toda sua informação hereditária que está contida em seu DNA (ou para alguns vírus, no RNA), incluindo tanto os genes como as regiões não-codificantes.

Mais precisamente, o genoma de um organismo é a seqüência completa do DNA de um conjunto de cromossomos. Por exemplo, um dos dois conjuntos de cromossomos para organismos diplóides (que têm cromossomos aos pares).

O termo genoma pode ser aplicado para o conjunto do DNA encontrado no núcleo da célula mas também pode ser aplicado a organelas que contêm seu próprio DNA, como o genoma mitocondrial ou dos cloroplastos.

Entretanto, quando se diz, por exemplo, que o genoma humano foi seqüenciado, em geral se diz que cada um dos cromossomos somáticos (não sexuais) e cada um dos dois cromossomos sexuais foi seqüenciado, cobrindo assim o material genético de ambos os sexos.

Homologia – Quando se diz que seqüências, proteínas ou genes são homólogos, se quer dizer que eles compartilham um ancestral comum, ou seja, são evolutivamente relacionados. Entidades relacionadas por homologia podem ser de dois tipos: ortólogas ou parálogas.

Nucleotídeo – Os nucleotídeos são compostos químicos formados por uma base nitrogenada, uma pentose e um grupo fosfato. Eles são os componentes básicos do DNA e RNA e também de outras moléculas chamadas cofatores. Os que compõem o DNA e RNA se diferenciam por sua base nitrogenada que no DNA pode ser A, C, G e T e A, C, G e U no RNA.

ORF – Uma ORF (“open reading frame”) corresponde ao trecho de DNA que pode ser traduzido em uma proteína ou em RNA. Começa com um codon de iniciação "ATG" e termina com um dos três códons de terminação ("TAA", "TAG" ou "TGA").

Ortólogos – Genes são ditos ortólogos quando são derivados de um ancestral comum mas estão presentes em organismos de espécies diferentes, ou seja se diz que eles são gerados a partir de eventos de especiação. Genes ortólogos costumam ter sua seqüência, estrutura e funções similares.

Parálogos – Genes ditos parálogos são genes homólogos que ocorrem em organismos de uma mesma espécie. Parálogos são criados a partir da duplicação de um gene, o que às vezes permite que genes adotem funções especializadas.

Procarioto – são organismos unicelulares que não possuem membrana envolvendo o núcleo da célula nem outras organelas que possuem membrana como mitocôndrias ou cloroplastos. Deste modo seu DNA que geralmente é composto por apenas um cromossomo circular fica disperso no citoplasma da célula.

A maioria dos procariotos são bactérias e os termos freqüentemente são usados como sinônimos. Além disso vale ressaltar que em organismos procariotos os genes correspondem a maior parte do DNA da célula e logo são facilmente encontrados, ao contrário do que acontece com organismos mais complexos onde as áreas que codificam genes são mais raras, chegando a poucos porcento do genoma, e então muito mais difíceis de serem encontradas.

Proteína – Proteínas são compostos orgânicos complexos e de alta massa molecular que consistem de aminoácidos unidos por ligações peptídicas. Proteínas são essenciais para a estrutura e funcionamento de todas as células de todos os seres.

Diversas proteínas desempenham uma grande variedade de funções biológicas. Algumas proteínas são enzimas que catalisam reações químicas, outras desempenham papéis estruturais ou mecânicos. Outras funções desempenhadas por proteínas incluem ainda resposta imunológica, armazenamento e transporte de vários compostos.

Proteínas são uma classe de biomoléculas que juntamente com os polissacarídeos, lipídeos e ácidos nucléicos formam os constituintes primários de organismos biológicos. Proteínas são, essencialmente, polímeros compostos por uma seqüência específica de aminoácidos e os detalhes dessa seqüência são armazenados em um gene. Por meio do processo de transcrição e tradução, a célula lê a informação genética e a usa para construir a proteína. Em muitos casos, a proteína resultante é ainda alterada quimicamente, antes de se tornar funcional. Além disso, é muito comum que proteínas funcionem juntas de modo a desempenhar uma função particular e freqüentemente se associam fisicamente formando um outro composto.

RNA – O RNA (“ribonucleic acid” ou em português ácido ribonucléico, ARN) é um polímero de nucleotídeos, freqüentemente com cadeia simples, cuja composição é muito semelhante à do DNA, com a diferença de ter uma ribose no lugar da desoxirribose e de usar a base uracila (U) ao invés da timina (T). O RNA é transcrito do DNA e serve como um modelo para a tradução dos genes em proteínas (mRNA), transporta aminoácidos para os ribossomos para sintetizar proteínas (tRNA) e também faz parte da composição dos próprios ribossomos (rRNA).

Seqüência – Nesse trabalho quando refere-se a uma “seqüência”, seja de DNA, RNA ou de uma proteína, refere-se ao que também é chamado de estrutura primária. A estrutura primária de uma molécula é a especificação atômica de sua composição, o que no caso de seqüências de DNA ou RNA é o equivalente a seqüência de nucleotídeos que os compõe, representados pelos nomes das bases que os diferenciam (abreviados pelas letras A, C, G, T ou U). Já a estrutura primária de uma proteína é dada pela seqüência de seus aminoácidos, que são de 20 diferentes tipos, cada um representado por uma letra.

Seqüenciamento – Seqüenciamento é o processo bioquímico pelo qual se determina a estrutura primária de uma molécula (em geral de DNA, RNA ou de uma proteína). Esse processo freqüentemente contém erros e medidas de qualidade (valor PHRED) são utilizadas para determinar a qualidade do resultado de um determinado seqüenciamento (Ewing et al, 1998). Deste modo, para que uma seqüência possa ser depositada em bancos de seqüências, ela deve ter um nível de qualidade acima de um valor determinado.

Via Bioquímica – Uma via metabólica ou bioquímica é uma série de reações químicas que ocorrem numa célula, catalizada por enzimas (proteínas) e que resulta na formação de um produto metabólico usado ou armazenado na célula ou no início de uma outra via metabólica. Muitas vias são complexas e envolvem vários passos na modificação da substância inicial até a formação do produto com as propriedades desejadas.