#VemPraFamart
famart.edu.br

CAPACITAÇÃO PROFISSIONAL

BIOINFORMÁTICA

famart
GRUPO EDUCACIONAL

BIOINFORMÁTICA

DÚVIDAS E ORIENTAÇÕES

Segunda a Sexta das 09:00 as 18:00

ATENDIMENTO AO ALUNO

editorafamart@famart.edu.br

famart
GRUPO EDUCACIONAL

Sumário

PRINCÍPIOS BÁSICOS DE BIOLOGIA MOLECULAR

A biologia molecular é uma ciência em constante evolução. Inicialmente, os estudos nessa área focaram a complexidade estrutural e a funcionalidade dos ácidos nucleicos nas mais diversas espécies. Nos últimos anos, porém, investimentos no desenvolvimento das técnicas de biologia molecular, levaram ao aprimoramento do sequenciamento gênico e da identificação de proteínas. Um exemplo é a finalização do sequenciamento do genoma humano realizado na última década, o qual só foi possível devido ao aperfeiçoamento dos sequenciadores automáticos.

A sofisticação técnica trouxe consigo uma grande quantidade de dados para serem analisados e corretamente interpretados. De tal modo, bancos de dados genéticos públicos contendo bilhões de dados moleculares e programas de análises moleculares começaram a ser desenvolvidos com o intuito de integrar o maior número possível de achados científicos, e com isso facilitar o trabalho dos pesquisadores.

A grande questão dessas bases de dados e ferramentas de análise, é que elas requerem programas analíticos avançados para catalogar e representar adequadamente as informações ali depositadas, e nem sempre os cientistas acostumados com a rotina laboratorial têm familiaridade com os processos matemáticos e computacionais requeridos por essas ferramentas.

Nesse contexto, tornou-se clara a necessidade de uma disciplina capaz de integrar conhecimentos de informática, estatística e biologia molecular, visando o aprimoramento profissional, bem como o melhor entendimento dos inúmeros dados gerados diariamente pelas pesquisas científicas. Essa disciplina é a tão comentada bioinformática.

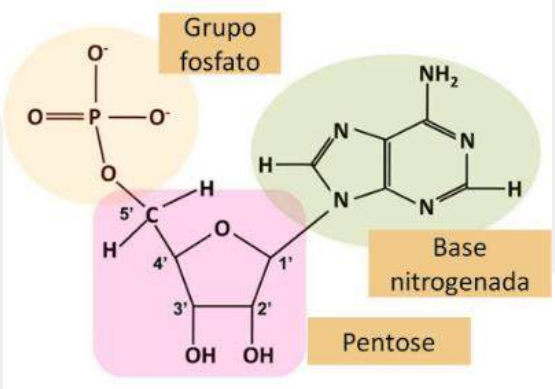
Em 1865, Mendel mostrou que existiam “fatores” no nosso organismo responsável pela transmissão da informação genética. Hoje sabemos que esses “fatores” nada mais são do que os genes, constituídos basicamente dos ácidos nucleicos, ácido desoxirribonucleico (DNA) e ácido ribonucleico (RNA).

ESTRUTURA DO ÁCIDO DESOXIRRIBONUCLEICO (DNA)

O ácido desoxirribonucleico (DNA) é uma molécula longa, composta de nucleotídeos, os quais são formados por uma pentose (desoxirribose), uma base nitrogenada (adenina, timina, guanina ou citosina) e um grupamento fosfato. A principal função do DNA é o armazenamento de informações genéticas para a posterior síntese de RNAs e proteínas, sendo o responsável pela transmissão das características genéticas entre os seres vivos, tais como cor dos olhos, pele, entre outras.

Um aspecto importante quando consideramos a estrutura do DNA e mais precisamente da estrutura dos nucleotídeos que o compõem, é a orientação da ligação entre os três componentes que compõem esses nucleotídeos, a qual é fundamental para a replicação correta da molécula de DNA. A ligação entre a base nitrogenada e a desoxirribose (açúcar) ocorre por meio de uma ligação covalente, enquanto que a ligação da desoxirribose com o grupamento fosfato ocorre por uma ligação fosfoéster.

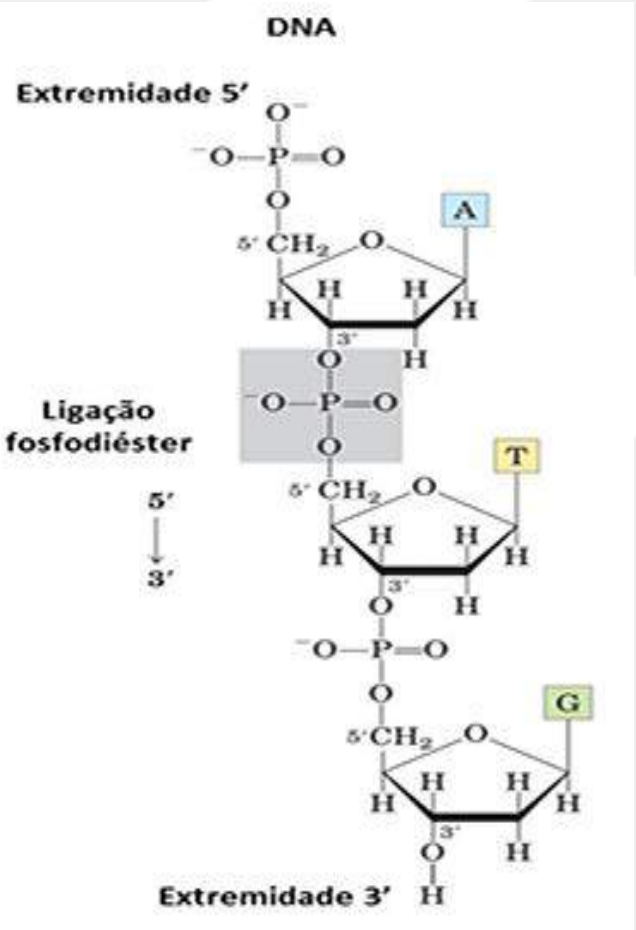
ESTRUTURA DE UM NUCLEOTÍDEO QUE COMPÕE A MOLÉCULA DE DNA



Para que a molécula de DNA seja formada, os nucleotídeos se ligam por ligações fosfodiéster. Dessa forma, a hidroxila do carbono 3 da desoxirribose do primeiro nucleotídeo se liga ao grupamento fosfato ligado a hidroxila do carbono 5 da desoxirribose do segundo nucleotídeo. Devido a essa formação, a molécula de DNA apresenta uma direção determinada, ou seja, em uma extremidade da cadeia

temos uma hidroxila livre no carbono 5 da primeira pentose e na outra extremidade temos livre a hidroxila do carbono 3 da última pentose. Esse dado é fundamental, pois determina que a replicação do DNA ocorra sempre no sentido 5´para 3´.

FIGURA: DEMONSTRAÇÃO DAS LIGAÇÕES FOSFODIÉSTER ENTRE NUCLEOTÍDEOS, MOSTRANDO O SENTIDO DE FORMAÇÃO DA MOLÉCULA DE DNA

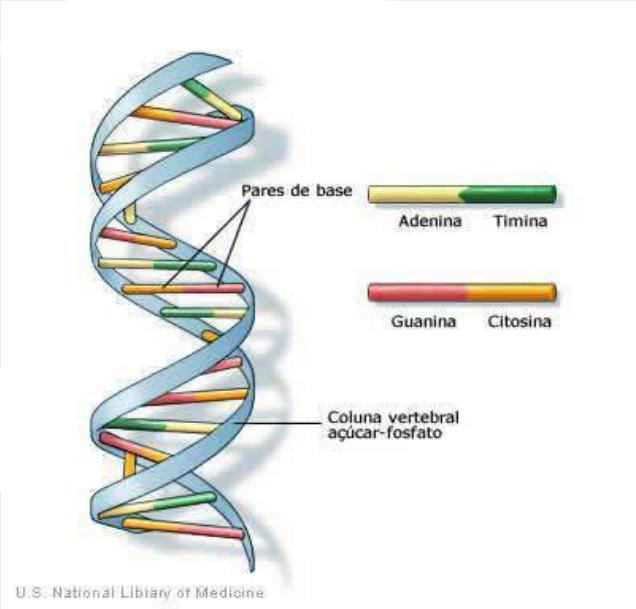


<http://medicina.med.up.pt

Um grande avanço nos estudos sobre a estrutura do DNA ocorreu em 1953, quando Watson e Crick por meio de estudos de difração de raios X propuseram um modelo tridimensional para a estrutura do DNA. Os pesquisadores mostraram que a molécula de DNA é uma dupla hélice com giro para direita em que duas cadeias polinucleotídeas são helicoidizadas uma ao redor da outra. Cada cadeia polinucleotídea consiste de uma sequência de nucleotídeos unidos por ligações

fosfodiéster, sendo os dois filamentos polinucleotídeos mantidos em sua configuração helicoidal por pontes de hidrogênio. As bases nitrogenadas estão no interior dessa estrutura tridimensional, enquanto que a pentose e o grupamento fosfato se dispõem externamente à molécula.

FIGURA: REPRESENTAÇÃO ESQUEMÁTICA DA DUPLA HÉLICE DE DNA



<http://curiofisica.com.br/

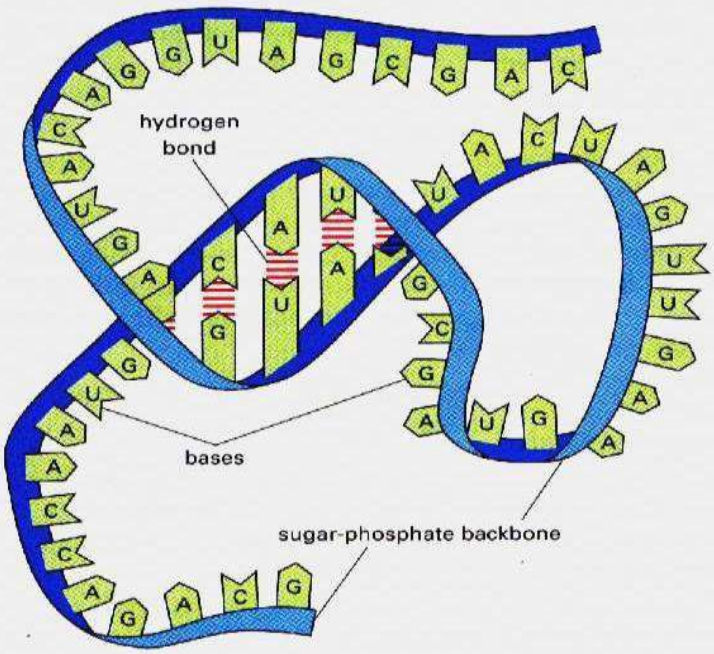
Entre as características básicas da molécula de DNA destaca-se a complementaridade das bases (adenina pareia com timina, e citosina pareia com guanina), a qual torna o DNA adequado para estocar e transmitir a informação genética de geração para geração; e a polaridade oposta das fitas de DNA, ou seja, que uma fita tem a direção da síntese (5´ para 3´) enquanto que a outra está em sentido oposto (3´para 5´). Essa disposição antiparalela das fitas é fundamental para entendermos como ocorre a replicação do DNA posteriormente.

ESTRUTURA DO ÁCIDO RIBONUCLEICO (RNA)

O ácido ribonucleico (RNA) é formado a partir de um molde de DNA, por um processo denominado de transcrição gênica. O RNA é uma molécula de fita

simples, formado por nucleotídeos compostos de uma pentose (ribose), uma base nitrogenada (adenina, uracila, guanina e citosina) e um grupamento fosfato.

FIGURA: REPRESENTAÇÃO ESQUEMÁTICA DE UMA MOLÉCULA DE RNA



<http://www.uic.edu/classes/phys/phys461/phys450/ANJUM04/>.

Vários tipos de moléculas de RNA foram identificados, das quais três se destacam: RNA mensageiro (RNAm), RNA ribossômico (RNAr) e RNA transportador (RNAt).

Os RNAm são moléculas de RNA traduzidas nos ribossomos, ou seja, são os RNAs que contêm a informação genética para a codificação das proteínas. Esse tipo de RNA compreende apenas 5% do RNA total da célula. Os RNAm dos organismos eucarióticos apresentam algumas características estruturais importantes, tais como a presença de uma cauda de adenina na extremidade 3´ (cauda poli-A) e uma estrutura denominada de cap na extremidade 5´ da cadeia de RNA. Por outro lado, os RNAt são pequenas moléculas de RNA que funcionam como adaptadores, transportando o aminoácido específico ao sítio de síntese de proteínas. Compreendem aproximadamente 15% do RNA total da célula. Os RNAr são

componentes tanto estruturais como catalíticos dos ribossomos, local onde ocorre o processo de tradução propriamente dito. No citoplasma dos organismos eucarióticos, encontramos quatro RNAr de tamanhos diferenciados (5S, 5,8S, 18S e 28S). Juntos compreendem 80% do RNA total celular.

Há ainda outros tipos de RNAs, tais como os pequenos RNA nucleares que compõem os spliceossomos, estruturas fundamentais para a retirada dos íntrons durante a transcrição gênica; e os microRNAs que são filamentos de cerca de 22 nucleotídeos, capazes de bloquear a expressão gênica de RNAm complementares ou parcialmente complementares, levando a degradação desses RNAm ou reprimindo o processo de síntese proteica. No entanto, o real papel desses RNAs nos diversos mecanismos moleculares ainda está sendo estudado.

REPLICAÇÃO DO DNA

A velocidade de síntese de um novo filamento de DNA em humanos é de cerca de 3.000 nucleotídeos por minuto, sendo que em bactérias cerca de 30.000 nucleotídeos são adicionados a uma cadeia de DNA nascente por minuto. Diante desses fatos, fica claro que a maquinaria de replicação do DNA deve funcionar de forma rápida e precisa. Estima-se que ocorra um erro por bilhão de nucleotídeos incorporados após a síntese e correção dos erros durante e logo após o processo de replicação do DNA.

A replicação do DNA ocorre de forma semiconservativa, visto que cada um dos filamentos complementares da dupla fita parental é conservado durante o processo. Além disso, a replicação do DNA é iniciada em origens únicas, as quais são caracterizadas por uma sequência de bases específicas denominadas origem da replicação. É interessante ressaltar que a partir da origem de replicação, esse processo ocorre bidirecionalmente.

Dessa forma, uma nova fita de DNA é sintetizada na direção 5´ para 3´. No entanto, considerando o sentido da replicação e a natureza antiparalela do DNA, percebemos que a fita utilizada como molde só pode ser lida da extremidade 3´ para 5. Como a abertura da dupla fita é gradual a partir da forquilha de replicação e o

sentido de replicação deve ser obrigatoriamente na direção 5´ para 3´, a síntese das duas fitas ocorre de forma descontínua, ou seja, em sentidos opostos. Sendo assim, uma fita é sintetizada continuamente, sendo denominada de fita líder, enquanto que a outra (fita tardia) é sintetizada a partir de pequenos fragmentos denominados de fragmentos de Okasaki.

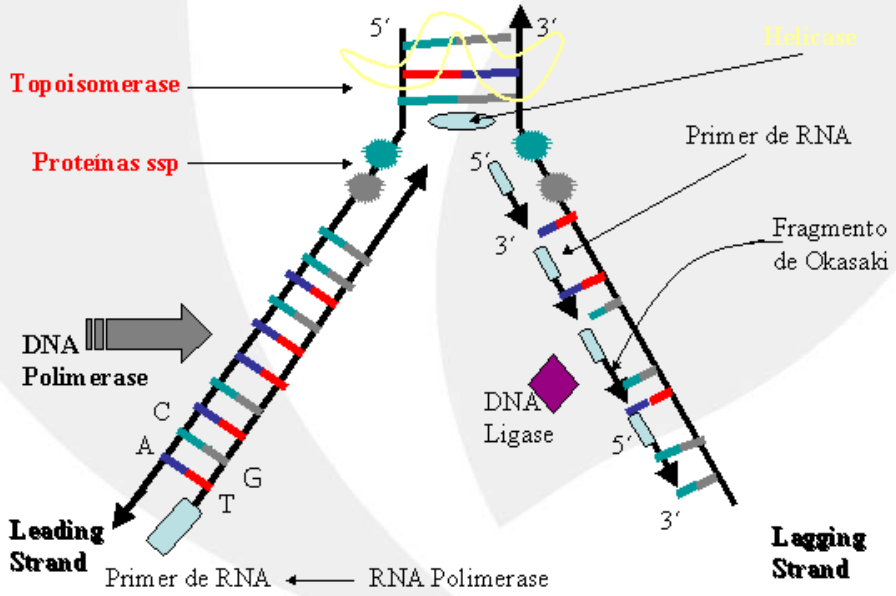
Didaticamente, o processo de replicação do DNA pode ser dividido em três etapas: iniciação, alongamento e terminação. A iniciação é considerada a única fase regulada da replicação do DNA, e visa assegurar que o DNA seja replicado somente uma vez por ciclo celular. Como o próprio nome diz, essa fase corresponde ao início da replicação do DNA, na qual as enzimas helicases agem separando a dupla fita de DNA. Seguida a essa etapa, inicia a fase de alongamento, a qual corresponde à síntese das fitas líder e tardia propriamente dita.

A síntese da fita líder inicia-se com a síntese de um pequeno segmento de RNA (primer) pela enzima primase, visto que a DNA polimerase, enzima chave desse processo, não tem a capacidade de fazê-lo pela ausência de grupos hidroxila (OH) expostos. Após a síntese do primer, a enzima DNA polimerase está apta a adicionar os desoxirribonucleotídeos, sintetizando assim a nova molécula de DNA. A síntese da fita líder prossegue continuamente até o final da molécula, na qual o primer é então retirado e substituído por DNA.

Por outro lado, a síntese da fita tardia ocorre pela formação de pequenos fragmentos. Inicialmente, a exemplo da síntese da fita líder, também é sintetizado um primer de RNA, e a enzima DNA polimerase adiciona desoxirribonucleotídeos à cadeia de DNA nascente. A grande complexidade dessa síntese reside no fato de que apenas uma DNA polimerase deve sintetizar a fita líder e tardia ao mesmo tempo. Para que isso ocorra, uma pequena alça na fita tardia é criada, a qual visa aproximar as duas forquilhas de replicação. Como a fita tardia é sintetizada descontinuamente, a síntese do primer de RNA é repetida diversas vezes durante a replicação do DNA. No final do processo, todos os primers são retirados por enzimas específicas e os fragmentos de Okasaki unidos, concluindo com isso a síntese da fita tardia do DNA.

A última etapa da replicação do DNA é a terminação. Considerando um modelo de organismo procariótico, tal como a Escherichia coli, podemos inferir que como seu DNA é disposto de forma circular e a replicação ocorre bidirecionalmente, chegará um ponto em que as duas forquilhas de replicação se encontrarão e terminarão a síntese daquela nova molécula de DNA. Em eucariotos, porém, a questão é mais complexa. Na fita líder, a replicação ocorre normalmente até o final do molde parental. Na fita tardia, porém, como ela está sendo sintetizada em sentido oposto, a extremidade da fita torna-se difícil de replicar. A solução envolve a ação das enzimas telomerases que sintetizam os telômeros, os quais são compostos de várias cópias de uma sequência consenso, assegurando assim que a fita tardia seja replicada adequadamente até o término do processo de replicação.

FIGURA: REPRESENTAÇÃO ESQUEMÁTICA DA REPLICAÇÃO DO DNA, MOSTRANDO A AÇÃO DAS DIVERSAS ENZIMAS ENVOLVIDAS NESSE PROCESSO



Leading strand corresponde a fita líder, e lagging strand corresponde a fita tardia.http://www.sistemanervoso.com

TRANSCRIÇÃO

A transcrição é o processo pelo qual, a partir de uma molécula de DNA, é sintetizada uma molécula de RNA. Por meio desse processo, todos os tipos de RNAs são formados (RNAm,

RNAr, RNAt, entre outros). Há diversas semelhanças entre a síntese de DNA e a síntese de RNA, porém a função biológica de cada um desses eventos é diferente. A síntese de DNA é um processo preciso e uniforme, pois o objetivo é a formação de uma nova molécula de DNA idêntica à molécula original. Por outro lado, a transcrição é um processo bastante variável e seletivo que espelha a condição fisiológica da célula, a qual pode requerer a expressão ou não de determinado gene em determinado momento.

De modo geral, a transcrição ocorre a partir da informação contida na sequência de nucleotídeos de uma molécula de DNA dupla fita, ocorrendo no sentido 5´ para 3. Nesse processo, apenas uma das fitas de DNA (fita molde) é utilizada para a síntese do RNAm, a qual segue as mesmas regras de antiparalelismo e complementaridade, com exceção do pareamento de uma uracila com adenina, ao invés de uma timina. Dessa forma, o RNAm recém-sintetizado (5´ para 3´) é complementar a fita que serviu de molde (3´ para 5´) e idêntico à outra fita de DNA (5´ para 3´).

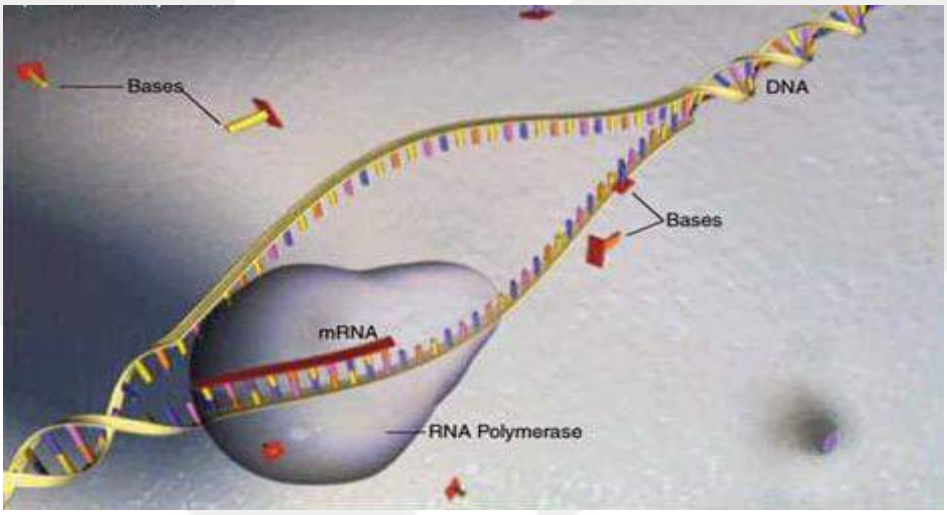
De forma semelhante ao que ocorre com a replicação do DNA, o processo de transcrição gênica também pode ser didaticamente dividido em três etapas: iniciação, alongamento e terminação. A fase de iniciação da transcrição requer regiões regulatórias, acoplamento de fatores de transcrição e fatores da RNA polimerase, enzima chave no processo de transcrição gênica. As regiões regulatórias da transcrição apresentam em comum uma região promotora, a qual é composta do sítio de início da transcrição, uma região conservada TATA box, uma região regulatória usptream CAAT box, e algumas outras sequências upstream que aumentam a eficiência de reconhecimento do promotor, e elementos enhancers que visam aumentar a expressão gênica.

O início da transcrição envolve uma cascata, na qual diversos fatores de transcrição se ligam ao DNA para iniciar o processo. Depois de montado o complexo

proteico nas regiões regulatórias, a mesma está apta a receber a enzima RNA polimerase e formar o complexo de iniciação da transcrição. Essa enzima abre a dupla fita de DNA e coloca a primeira base, formando a bolha de transcrição, estrutura característica do processo.

Depois de iniciada a transcrição gênica, a cadeia de RNA nascente é alongada. Nesse processo de alongamento, outros fatores acessórios são necessários. A bolha de transcrição se move pela fita de DNA, abrindo a dupla fita de DNA à frente e fechando atrás de si. Os erros que porventura possam ocorrer na sequência da cadeia de RNA nascente são corrigidos pela ação da RNA polimerase durante essa etapa. Após o alongamento da cadeia de RNA, ocorre o término desse processo de transcrição gênica. Sabe-se que existem regiões consenso de DNAs ricas em timina, as quais são regiões sinais codificadas pela molécula de DNA que indicam onde deve ser a extremidade 3´. Enzimas que se ligam ao RNA acabam por reconhecer esses sinais e clivam o fragmento de RNA nascente nessas regiões.

FIGURA: REPRESENTAÇÃO ESQUEMÁTICA DA TRANSCRIÇÃO DO DNA EM RNA



A figura destaca a bolha de transcrição e a ação da RNA polimerase durante esse processo. <http://www.odnavaiaescola.com.br/transcricao.html>.

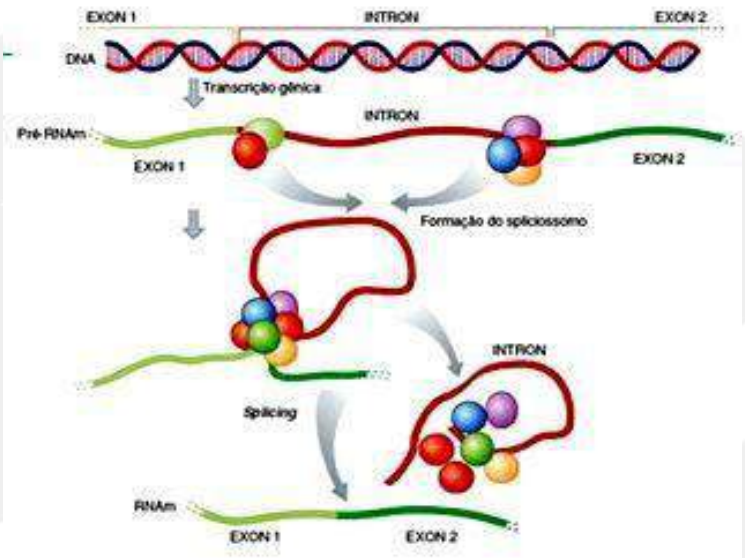
PROCESSAMENTO DO RNA

Os diversos RNAs formados durante o processo de transcrição são denominados de transcritos primários, os quais não representam a molécula de RNA madura, ou seja, àquela cuja sequência e estrutura correspondem ao RNA funcional.

Diante disso, esses transcritos primários precisam sofrer modificações, a fim de se tornarem RNAs maduros. Os mecanismos de processamento do RNA são bastante variados e dependem do tipo de RNA sintetizado. Por exemplo, as modificações do RNAm incluem a adição do cap na extremidade 5´, o qual protege a extremidade 5´ do transcrito da ação de enzimas exonucleases, além de facilitar o splicing do RNA e o transporte do núcleo para o citoplasma; a adição de uma cauda poli A na extremidade 3´ que facilita o transporte do RNAm para o citoplasma, estabiliza a molécula no citoplasma, podendo também estar relacionada com o reconhecimento da molécula de RNAm pelo complexo que posteriormente realizará a tradução; a metilação de regiões específicas do RNAm que podem ajudar a proteger as porções do transcrito primário que precisam de preservação; e os comentados mecanismos de splicing alternativo.

O splicing nada mais é do que a excisão de íntrons do pré-transcrito. De forma geral, o transcrito primário é composto de dois tipos de sequências, os éxons, os quais são as porções codificadoras que estarão presentes no RNA maduro, bem como os íntrons, que são sequências que não estarão presentes no RNA maduro, e, portanto, não codificam proteínas. O splicing não ocorre de forma aleatória, existem sequências consenso presentes nos íntrons que sinalizam o local onde deve ocorrer o splicing, e são fundamentais para a excisão correta dos elementos intrônicos. Esse processo envolve a quebra de ligações fosfodiéster na junção éxon-íntron, e a formação de outra ligação entre as extremidades do éxon. Comumente esse processo é mediado por spliceossomos, que são estruturas compostas de um grande número de ribonucleoproteínas e outras proteínas acessórias, desfeito após a remoção dos íntrons. No entanto, o splicing também pode ocorrer por um mecanismo denominado de “auto- splicing”, na qual a excisão intrônica ocorre pela atividade catalítica do próprio RNA precursor.

FIGURA: REPRESENTAÇÃO ESQUEMÁTICA DO PROCESSO DE SPLICING ALTERNATIVO, MOSTRANDO A FORMAÇÃO DO SPLICEOSSOMO E A REMOÇÃO DAS SEQUÊNCIAS INTRÔNICAS



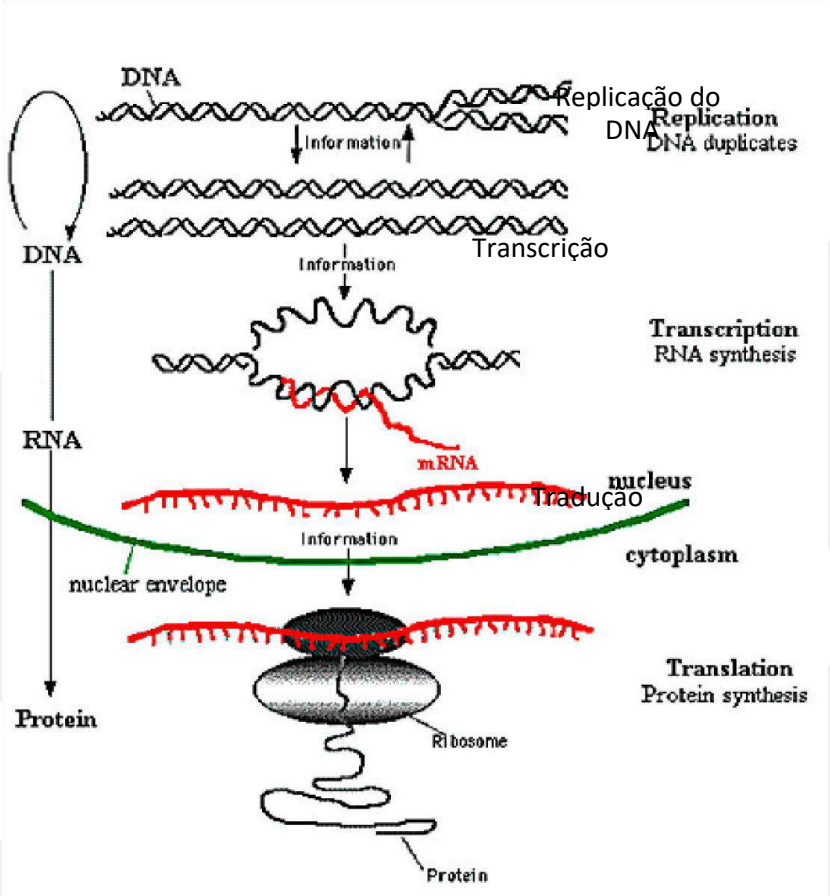
http://www.algosobre.com.br/biologia/o-que-e-splicing.html

TRADUÇÃO

As proteínas são as moléculas mais abundantes e diversas dos sistemas biológicos. São compostas de polipeptídeos, os quais são formados por aminoácidos. Sabemos que existem 20 aminoácidos diferentes para formar os vários tipos de proteínas identificadas em mamíferos.

A informação genética está armazenada nos cromossomos e é transmitida às células filhas pela replicação do DNA, expressa por meio da transcrição em RNA, e no caso específico do RNAm traduzido em uma proteína. Esse fluxo de informação genética, do DNA para a proteína é conhecido como dogma central da biologia molecular.

FIGURA: DOGMA CENTRAL DA BIOLOGIA MOLECULAR



Replicação do DNA

Transcrição

Tradução

<http://www.biomedicina3l.com

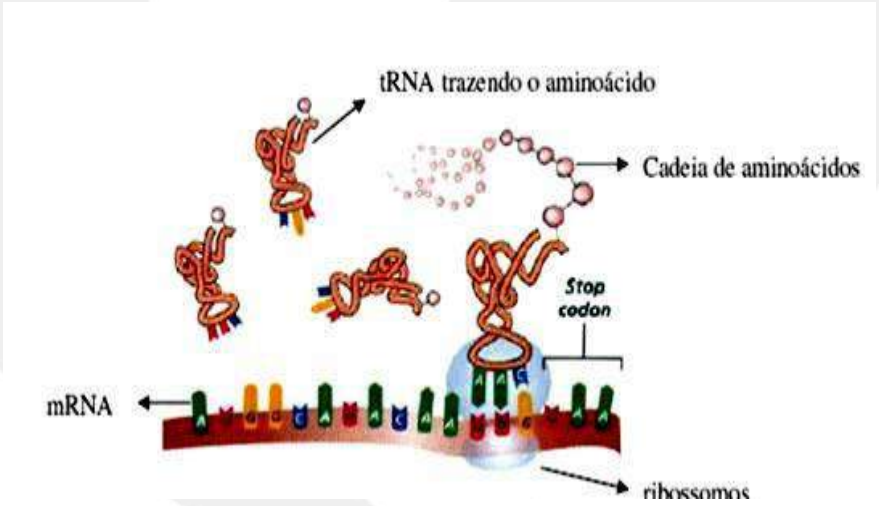
Antes de descrever o processo de tradução propriamente dito, vale algumas considerações sobre o código genético. Esse código nada mais é do que a forma pelo qual o DNA é estocado no nosso organismo. Esse código fornece uma correspondência entre as sequências de bases nucleotídicas e uma sequência de aminoácidos. O código genético apresenta três características fundamentais que o definem: especificidade, universalidade e redundância, visto que um determinado aminoácido pode ser especificado por mais de um códon, mas cada códon só pode designar um aminoácido.

Diferente dos outros processos citados, a síntese proteica ocorre no citoplasma das células. Para que o processo de tradução ocorra é necessário um grande número de componentes, os quais incluem todos os aminoácidos encontrados no produto proteico final, o RNAm a ser traduzido, RNAt, ribossomos

funcionais, fontes de energias e fatores proteicos necessários à iniciação, alongamento e terminação. É importante ressaltar que o RNAm não se liga diretamente aos aminoácidos, mas sim, o RNAt, o qual fornece a ligação molecular entre a sequência de bases do RNAm e a sequência de aminoácidos da proteína.

A iniciação da síntese proteica envolve a reunião de uma série de componentes, os quais incluem duas subunidades ribossômicas, o RNAm a ser traduzido, o aminoacil-RNAt para metionina que é o primeiro aminoácido a ser incorporado no peptídeo nascente, GTP que fornece energia para que o processo ocorra, e fatores de iniciação que facilitam a montagem desse complexo para síntese proteica. A elongação envolve a adição de aminoácidos à extremidade carboxila da cadeia polipeptídica em formação. A formação das ligações peptídicas entre os aminoácidos é catalisada por enzimas peptidiltransferases. Após a formação da ligação peptídica, o ribossomo avança três nucleotídeos em direção a região 3´ terminal do RNAm. Isto causa a liberação do RNAt descarregado. O término do processo ocorre quando um dos três códons de encerramento move-se para o sítio de síntese.

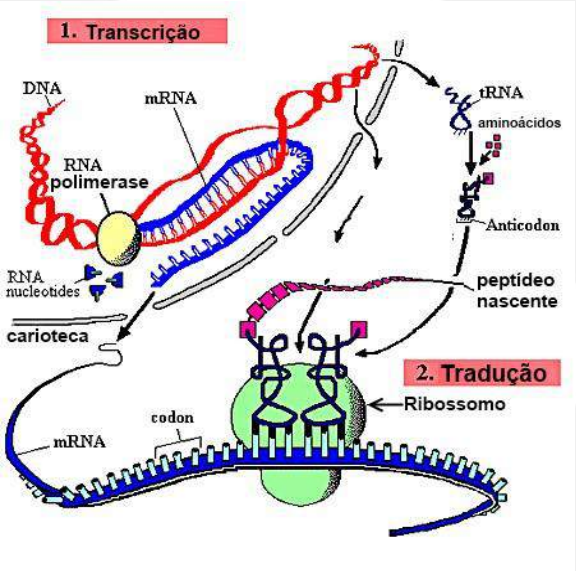
FIGURA: REPRESENTAÇÃO ESQUEMÁTICA DO PROCESSO DE TRADUÇÃO



<http://www.enq.ufsc.br

Depois de concluídas todas as etapas citadas, as proteínas sofrem ainda modificações pós-traducionais, as quais incluem adição de cadeias laterais de carboidratos aos polipeptídeos, modificações químicas, clivagens em unidades

polipeptídicas menores, combinação com outros polipeptídeos para formar proteínas maiores, entre outras. Todas essas modificações têm como objetivo tornar a proteína funcional, sendo necessárias, por exemplo, para produzir um dobramento apropriado da proteína final. A Figura traz um resumo dos processos de replicação, transcrição e tradução apresentados.



TÉCNICAS BÁSICAS DE BIOLOGIA MOLECULAR E BIOINFORMÁTICA

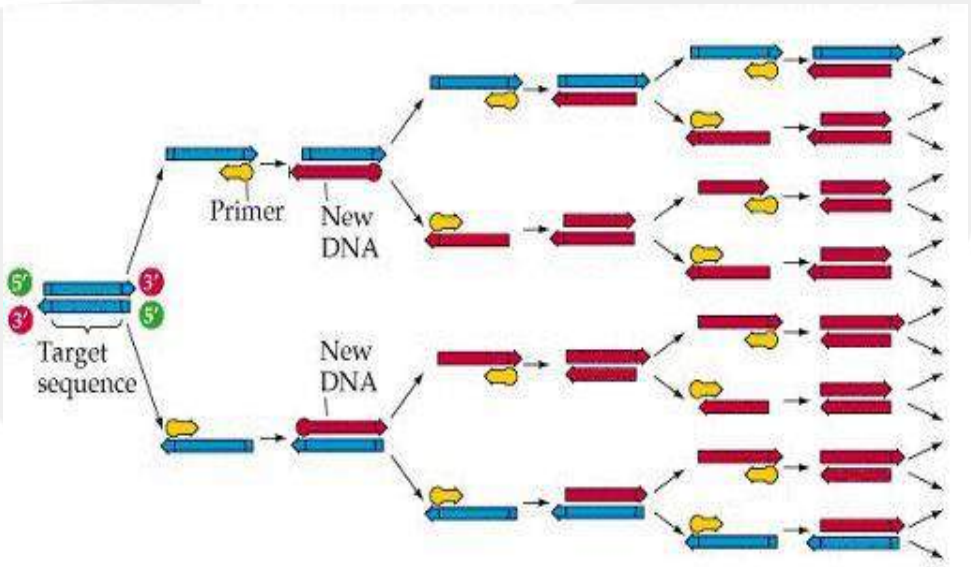
As técnicas de biologia molecular estão em constante evolução. Procedimentos técnicos adequados para os mais diversos tipos de estudos estão disponíveis.

Muito do que sabemos sobre a estrutura gênica foi obtida de estudos moleculares de genes e cromossomos por meio da tecnologia de DNA recombinante. Nesse contexto, sabemos que o primeiro passo para termos uma molécula de DNA recombinante é a clonagem de genes específicos, um processo bastante complexo que envolve o isolamento do gene de interesse, seguido da inserção em um elemento genético autorreplicante, tais como plasmídeos ou cromossomos virais, e a amplificação do gene de interesse durante a replicação do cromossomo viral ou plasmidial em um hospedeiro apropriado. Após o gene ter sido clonado, o mesmo pode ser manipulado de diversas formas, permitindo que investigações sobre a estrutura, bem como sobre a função gênica sejam realizadas.

Comumente um gene clonado é sequenciado, ou seja, é determinada a sequência de nucleotídeos que compõem esse gene.

Técnicas de extração de DNA e RNA, as quais visam extrair esses ácidos nucleicos dos mais diversos tecidos biológicos, por meio de técnicas como salting out, fenol-clorofórmio, entre outras para extração de DNA, e método do Trizol para extração de RNA são utilizadas rotineiramente nos laboratórios de biologia molecular. A reação em cadeia da polimerase (PCR), proposta em 1985 por Saiki e colaboradores, permitiu um grande avanço no estudo da biologia molecular. Essa técnica tem por objetivo amplificar uma sequência específica de DNA por meio de uma enzima termoestável in vitro. O princípio desse método é o anelamento de oligonucleotídeos específicos em regiões específicas da molécula de DNA, visando amplificar um determinado fragmento por ação de uma enzima taq polimerase. Em cada ciclo as fitas servem de molde para a geração de novas fitas.

FIGURA: REPRESENTAÇÃO DA REAÇÃO EM CADEIA DA POLIMERASE (PCR)



Essa imagem mostra a sequência lavo (target sequence), o anelamento dos primers específicos e a subsequente formação de uma nova fita de DNA (new DNA).

FONTE: Purves, W.K. et al. Life Wire, 2004.

A PCR serve de partida para o desenvolvimento de inúmeras técnicas, tais como digestão enzimática para avaliação de mutações e polimorfismos que alteram sítios de restrição enzimáticos, sequenciamento automático direto para avaliação da sequência de um gene ou recentemente de um genoma inteiro, clonagem gênica, entre inúmeras outras. Um grande avanço na técnica de PCR foi o desenvolvimento da PCR em tempo real, que possibilita análises de expressão gênica e de discriminação alélica de forma mais fácil e precisa do que outros métodos anteriormente empregados.

Todas as técnicas anteriormente citadas se interpolam facilmente com as ferramentas de bioinformática na rotina científica e laboratorial. Para exemplificar esse fato, considere que para desenharmos primers específicos para determinado gene precisamos conhecer a sequência desse gene, a qual está disponível em bancos de dados gênicos na internet. Caso queiramos saber se uma variante altera ou não sítios crípticos de splicing, podemos utilizar programas genéticos disponíveis na rede; para saber se determinada variante altera ou não a estrutura proteica final podemos lançar mão de programas que avaliam as características físico-químicas dos diversos aminoácidos, e avaliam os efeitos dessa troca para a proteína. Ainda, avaliações da estrutura tridimensional de uma proteína também podem ser realizadas por análises computacionais, entre inúmeras outras análises que podem ser realizadas por meio das técnicas de bioinformática.

Recentemente, o desenvolvimento dos sequenciadores de nova geração, os quais apresentam uma tecnologia de sequenciamento de DNA que permite gerar informações sobre milhões de pares de bases em uma única corrida, está revolucionando a pesquisa genômica estrutural e funcional. Diante desse cenário científico, a bioinformática e os profissionais especialistas nessa área tornam-se ainda mais fundamentais, visto a enormidade de dados produzidos e a dificuldade de interpretação dos mesmos pelos pesquisadores. A bioinformática não é mais a ciência do futuro, mas sim, a necessidade do presente.

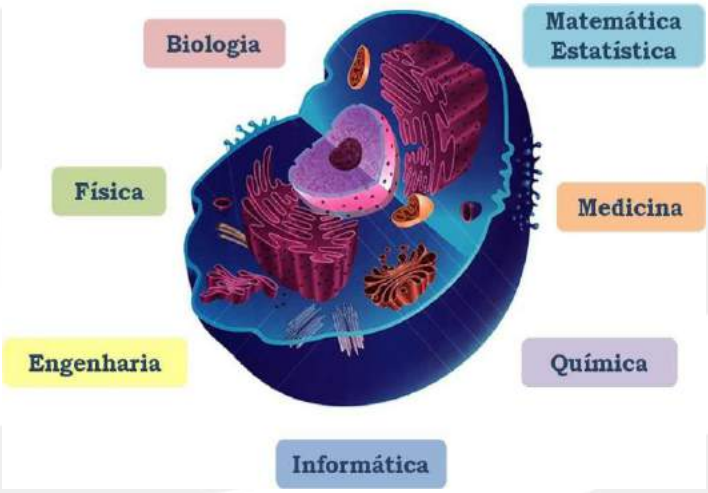
CONCEITOS DE BIOINFORMÁTICA

A bioinformática é comumente definida como a aplicação de técnicas computacionais para entender e organizar informações associadas com macromoléculas biológicas. Em 2000, Brown definiu bioinformática como o uso de computadores para aquisição, manejo e análise da informação biológica, sendo a bioinformática resultado da intersecção da biologia molecular, biologia computacional, medicina clínica, bases de dados de informática, internet e da análise de sequências.

Luscombe et al. ampliaram os conceitos básicos de bioinformática e a definiram como o ato de conceituar a biologia em termos moleculares e aplicar técnicas de informática, as quais derivam de disciplinas como matemática aplicada, ciências da computação e estatística, a fim de entender e organizar as informações associadas com essas moléculas em larga escala. Resumindo essa complexa definição, os autores conceituaram bioinformática como o manejo de sistemas de informação para biologia molecular, com inúmeras aplicações práticas no dia a dia dos pesquisadores. Segundo o NCBI (National Center for Biotechnology Information), a bioinformática é um campo da ciência, na qual diversas disciplinas confluem, tais como biologia, computação e tecnologia da informação. Para o NCBI, o objetivo da bioinformática é facilitar a busca por novos conhecimentos, e integrar os mesmos em bases de dados, visando com isso à formação de princípios universais no campo da biologia.

Finalmente, a definição de Bibgen parece integrar todos os outros conceitos anteriormente mencionados sobre a bioinformática. Para ele, a bioinformática é uma disciplina que se ocupa da aquisição, armazenamento, processamento, distribuição, análise e interpretação das informações biológicas, mediante a aplicação de técnicas e ferramentas procedentes da matemática, biologia e informática, com o propósito de compreender o significado biológico de uma grande variedade de dados.

FIGURA: REPRESENTAÇÃO ESQUEMÁTICA DAS PRINCIPAIS ÁREAS QUE INTEGRAM ABIOINFORMÁTICA



HISTÓRICO DA BIOINFORMÁTICA

Quando em 1953, Watson e Crick propuseram o modelo de dupla hélice do DNA, os pesquisadores não podiam imaginar o volume de informações que exponencialmente seriam geradas a partir daquele momento. Nas décadas que se seguiram ao trabalho de Watson e Crick, ferramentas computacionais, as quais eram inicialmente bastante simplificadas, começaram a ser desenvolvidas e possibilitaram a análise e a resolução de inúmeros questionamentos relacionados à estrutura do DNA, bem como à informação genética que codifica as proteínas, às propriedades estruturais dessas e os fatores que as regulam, assim como os eventos associados com a regulação genética, com as bases moleculares do desenvolvimento embrionário e com a evolução das vias metabólicas e bioquímicas.

Na década de 90, mais precisamente na segunda metade dessa década, começaram a surgir os sequenciadores automáticos, os quais geraram uma grande quantidade de sequências a serem armazenadas, exigindo com isso, o desenvolvimento de ferramentas computacionais mais eficientes. Diante dessa situação, a bioinformática iniciou o seu desenvolvimento. Filho et al. em uma revisão sobre o tema, afirmaram que os primeiros trabalhos desenvolvidos foram realizados

por profissionais da biologia e da informática. No entanto, as dificuldades de comunicação entre esses dois grupos de profissionais mostraram a necessidade de um novo profissional que tivesse conhecimentos suficientes para ligar essas duas áreas, o bioinformata.

Esse profissional deve ter conhecimentos para saber quais as questões biológicas relevantes e quais as melhores opções de abordagem computacional para as questões a serem resolvidas.

Historicamente, não podemos tratar da história da bioinformática sem fazer uma conexão com a história da biologia, visto que ambas se interconectam desde a origem. Desde os séculos XVIII e XIX, os pesquisadores enfrentaram problemas com o processamento das informações obtidas. O desenvolvimento da genética com a formulação das leis de Mendel há mais de cem anos e o descobrimento da estrutura do DNA em 1953, abriram um leque de investigações que culminaram no projeto genoma humano iniciado na década de 90 e concluído no início do século XXI. No entanto, desde a década de 60, o crescimento no número de sequências de aminoácidos conhecidas levou a aplicação pioneira dos computadores na biologia molecular. A partir desse período, a quantidade de dados que deveriam ser analisados cresceu consideravelmente, e os preços mais acessíveis dos computadores tornaram possível a introdução dos mesmos nos ambientes acadêmicos.

Margaret Dayhoff desenvolveu os primeiros programas para determinar a sequência de aminoácidos de uma proteína em 1965, e preparou o primeiro banco de dados de sequências de proteínas que evoluiu para o PIR (Protein Information Resource) em 1983. Os programas de comparação de sequências e de análise filogenética foram os primeiros avanços no campo da bioinformática na década de 60. Posteriormente, na década de 70, iniciaram-se as análises estruturais de macromoléculas.

No entanto, essas análises eram bastante limitadas devido à capacidade da informática disponível naquele tempo. Nesse mesmo período, os métodos de informática também começaram a ser aplicados no processamento das informações sobre os ácidos nucleicos. Programas para comparar sequências começaram a ser

desenvolvidos. O FASTA foi desenvolvido por volta de 1985, o Genbank no início da década de 1980, e o SwissProt por volta de 1987. No final dos anos 80, começou a se empregar o termo bioinformática para a ciência que integrava a informática e a biologia. A tabela ilustra os principais eventos relacionados à bioinformática até o projeto genoma humano.

TABELA: PRINCIPAIS EVENTOS HISTÓRICOS RELACIONADOS À BIOINFORMÁTICA

Ano	Principal Evento na Bioinformática
1962	Teoria da Evolução Molecular – Pauling
1965	Margaret Dayhoff – Atlas de seqüências de proteínas
1970	Algoritmo de Needleman-Wunsch – comparação entre seqüências
1977	Sequenciamento de DNA e desenvolvimento de softwares para análise de seqüências – R. Staden
1981	Desenvolvimento do algoritmo de Smith-Waterman
1982	Publicação do Release 3 do GenBank
1982	Sequenciamento do Genoma do Fago Lambda
1983	Algoritmo de busca de seqüências em bancos de dados - Wilbur-Lipman
1985	Comparação rápida de seqüências – FASTA
1988	Criação do *National Center for Biotechnology Information* (NCBI)
1988	Rede EMBnet para distribuição de bancos de dados
1990	Método mais rápido de comparação de seqüências –BLAST
1991	EST: Etiquetas de seqüência transcrita
1993	Criação do Sanger Center, Hinxton, UK
1994	Criação do EMBL (*European Bioinformatics Institute)*
1995	Sequenciamento completo dos primeiros genomas bacterianos
1996	Genoma completo da levedura S. cereviseae
1998	Genoma completo de C.elegans (multicelular)
1999	Genoma completo de D. melanogaster
2001	Genoma completo de Homo sapiens

Ainda no final da década de 80, programas de bioinformática mais avançados foram desenvolvidos nos centros acadêmicos e rapidamente se converteram em produtos comerciais, passando a ser distribuído como pacotes integrados de

ferramentas para a administração de dados de biologia molecular. A melhora nos sistemas computacionais permitiu um grande avanço das técnicas de aprendizado automáticas com clara aplicabilidade no campo da bioinformática. No final dos anos 90, a demanda por especialistas em bioinformática era notável, porém, poucas universidades ofereciam programas educativos para esse tema, os quais cresceram consideravelmente nos últimos anos. No exterior, podemos encontrar inúmeros cursos de formação em bioinformática, na sua grande maioria concentrados na América do Norte e Europa.

Em 1999, a ciência brasileira se destacou internacionalmente com o sequenciamento completo do DNA da bactéria Xylella fastidiosa. Esse trabalho contou com a utilização de softwares de sequenciamento genético com base na internet, correspondeu ao início da bioinformática no Brasil. Atualmente, diversos projetos de sequenciamento estão em curso no nosso país, brGene, OMM, PIGS, Leifsonia xyli, genoma do café, da banana, RioGene, entre outros.

FIGURA: SEQUENCIAMENTO DO GENOMA DA XYLELLA FASTIDIOSA

Umas das publicações de maior relevância da pesquisa brasileira, a qual foi capa da Nature. FONTE: Simpson, A.J.G. et al. Nature, 2000.

Em 2002 foi implantado o primeiro curso de especialização em bioinformática no Brasil pelo LNCC (Laboratório Nacional de Computação Científica), sendo que

nesse mesmo ano também foram autorizadas as criações de dois cursos de pós-graduação stricto sensu em bioinformática, um na USP e outro na UFMG, os quais estão em pleno funcionamento atualmente. Diante de tudo o que foi exposto, percebemos claramente que a bioinformática está crescendo consideravelmente, e cada vez mais será necessária para a interpretação dos dados em biologia molecular. Técnicas moleculares sofisticadas como microarray, sequenciamento de nova geração, entre outras, confirmam o papel decisivo da bioinformática para o entendimento dos bilhões de dados gerados por essas ferramentas inovadoras.

APLICAÇÕES DA BIOINFORMÁTICA

Todos os dias uma grande quantidade de informações é gerada pelo mapeamento gênico. Essas informações precisam ser decifradas, organizadas, decodificadas e armazenadas sistematicamente em bancos de dados computacionais, servindo de base para a realização de estudos médicos e biológicos. Esse armazenamento, bem como o processamento adequado dos dados gerados é realizado por meio da bioinformática. A bioinformática apresenta inúmeras aplicações práticas no dia a dia dos profissionais que trabalham diretamente com informações biológicas, tais como médicos, biólogos, biomédicos, farmacêuticos, agrônomos, entre outros.

Na medicina, por exemplo, a bioinformática pode auxiliar no diagnóstico e tratamento. Para exemplificar esse fato, podemos pensar na descoberta de uma mutação causal de uma doença. Para que essa mutação seja corretamente caracterizada e descoberta seu mecanismo patológico, é fundamental que saibamos a sequência da mesma e consigamos comparar com sequências disponíveis em bancos de dados biológicos.

Além disso, a partir da sequência de nucleotídeos, podemos saber qual a proteína codificada pelo gene em questão, e se a mutação altera a estrutura proteica final ou não. Ainda, podemos investigar mecanismos moleculares mais complexos, como alterações em regiões promotoras do gene ou mecanismos de splicing alternativos e regulação transcripcional, ou seja, podemos fazer uma gama de

análises moleculares por meio da bioinformática, visando o diagnóstico correto do paciente, bem como inferir sobre um possível tratamento.

Hoje, sabemos que pacientes com determinado genótipo podem ter um benefício melhor se tratados com determinada droga, enquanto que outros pacientes com genótipo diferente se beneficiam com o uso de outra droga. Todas essas inferências estão disponíveis em grandes bancos de dados na rede, os quais além de informar os clínicos sobre as recentes descobertas, também incentivam a realização de novos trabalhos científicos, a fim de expandirem os conhecimentos originais.

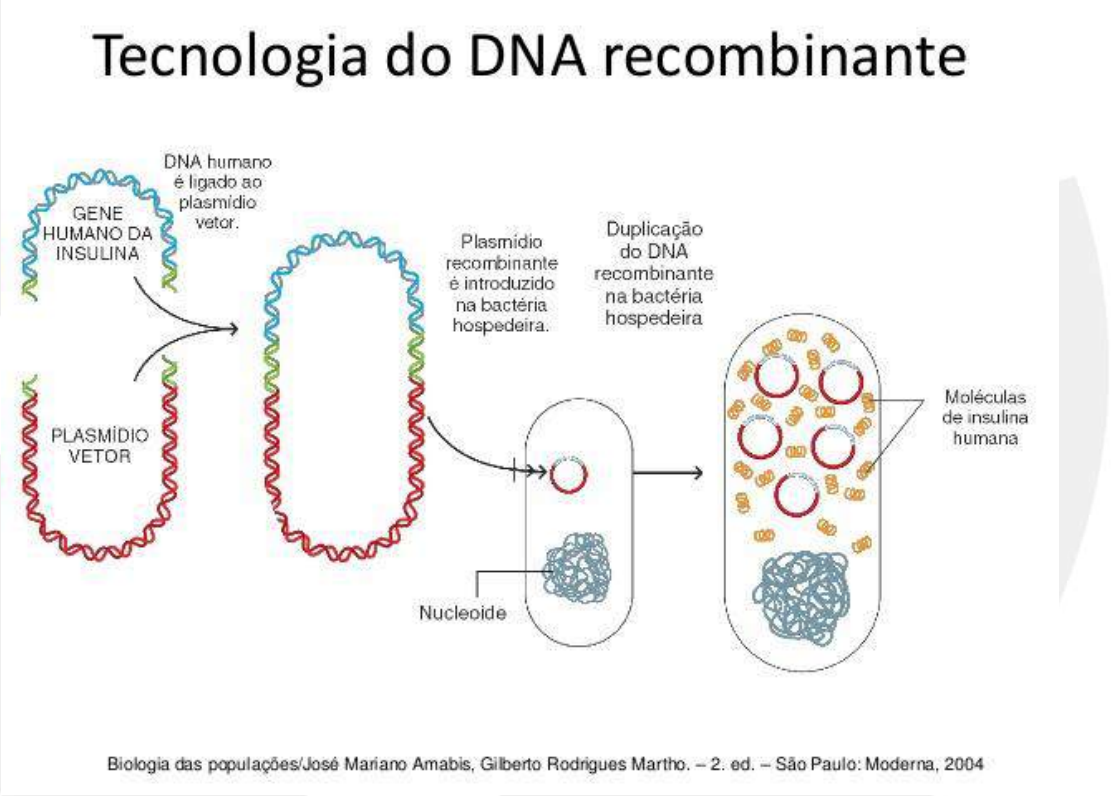
A bioinformática apresenta um papel central nas indústrias farmacêuticas. Para entendermos esse fato, precisamos considerar que o funcionamento adequado de uma terapia medicamentosa é produto da interação de várias moléculas, as quais quando em contato com o alvo biológico no organismo do paciente, inibem ou aceleram determinadas reações químicas, objetivando a atenuação do quadro patológico.

A formulação de um medicamento apresenta inúmeras etapas e ensaios tanto pré-clínicos como clínicos, os quais visam avaliar a toxicidade, eficácia, efeitos colaterais, entre outros aspectos. Para a produção de novas drogas é fundamental que sejam selecionadas moléculas biológicas com o perfil necessário para o uso daquele fármaco, e que as demais moléculas sejam descartadas, a fim de não interferirem com as moléculas selecionadas. A bioinformática é primordial na seleção dessas moléculas, orientando os pesquisadores na manipulação correta das mesmas. Ainda, programas disponíveis, sejam eles em nível de nucleotídeos ou de proteínas, permitem que se estudem alterações na estrutura gênica que codificam determinadas proteínas ou hormônios, de modo a criar fármacos alterados, visando com isso produzir drogas mais efetivas e de melhor qualidade.

Atualmente, a indústria farmacêutica conta com o apoio de centros de biotecnologia, a fim de melhorar suas pesquisas e o desenvolvimento de novos medicamentos. Diante das necessidades qualitativas e quantitativas, o setor farmacêutico parece ser o maior empreendedor da bioinformática. A figura mostra a tecnologia de DNA recombinante bastante utilizada pela indústria farmacêutica na

produção de novos fármacos. O desenvolvimento dessa tecnologia está intimamente ligado ao desenvolvimento da bioinformática, visto que precisamos de uma série de informações e testes gênicos e proteicos in silico (análises de bioinformática) para conseguirmos produzir um recombinante adequado e funcional, e principalmente para testar o resultado final desse recombinante nos organismos.

FIGURA: REPRESENTAÇÃO ESQUEMÁTICA DA TECNOLOGIA DE DNA RECOMBINANTE



Biologia das populações/José Mariano Amabis, Gilberto Rodrigues Martho. – 2. ed. – São Paulo: Moderna, 2004

<http://www.mundovestibular.com.br>.

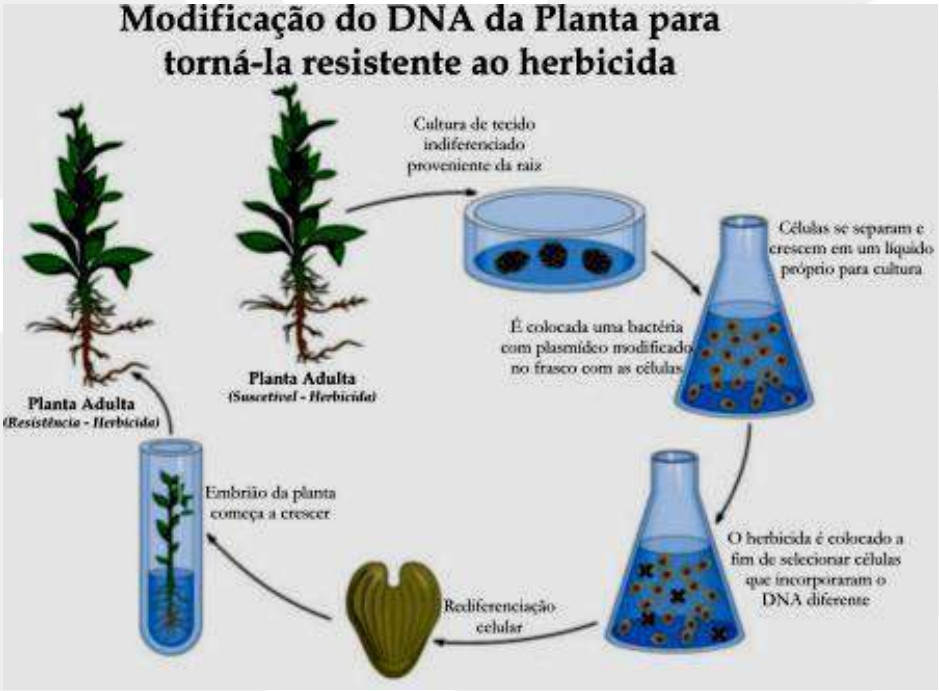
A tecnologia do DNA recombinante apresenta inúmeras aplicações, tais como na produção de vacinas sintéticas, na terapia gênica, na localização de genes relacionados a doenças, entre outras aplicações.

Na biotecnologia e na agricultura, as aplicações da bioinformática são semelhantes à anteriormente descrita, permitindo o desenvolvimento e produção de novos produtos, bem como o aumento da produtividade e melhoria dos alimentos, respectivamente. No Brasil, os estudos genéticos principalmente voltados para a

agropecuária alcançaram resultados satisfatórios, e utilizaram da bioinformática extensivamente para chegarem às suas conclusões finais. A figura traz um exemplo do uso tecnologia do DNA recombinante e da bioinformática para tornar uma planta resistente a um determinado herbicida.

A bioinformática é fundamental, por exemplo, para a identificação do gene de interesse, para a síntese do plasmídeo modificado, entre inúmeras outras aplicações durante o experimento.

FIGURA: REPRESENTAÇÃO ESQUEMÁTICA DA MODIFICAÇÃO GÊNICA PARA TORNAR UMA PLANTA RESISTENTE AO USO DE HERBICIDAS



<http://www.clinicaq.com.br>.

Diante de tudo o que foi exposto, percebemos as inúmeras aplicações da bioinformática nos mais diversos setores, e a sua grande contribuição para o desenvolvimento científico e tecnológico atual.

SISTEMAS OPERACIONAIS E LINGUAGENS DE PROGRAMAÇÃO

Antes de iniciarmos esse capítulo, convém ressaltar que o objetivo do mesmo é resumir algumas informações relevantes sobre sistemas operacionais e linguagem de programação em bioinformática. De modo algum, objetivamos explicar detalhadamente os componentes de um sistema operacional e as diversas linguagens de programação existentes.

O sistema computacional é formado um sistema físico (hardware) e um lógico (software). O sistema físico inclui o computador, monitor, CPU, placa de rede, disco rígido, entre outros componentes. Por outro lado, o sistema lógico inclui os aplicativos, como Microsoft Office, jogos, linguagens de programação, sistemas de bases de dados, sistemas operacionais, entre outros.

Segundo Filho et al., o sistema operacional é o principal programa de um computador, sendo responsável pelo gerenciamento da memória, acesso aos discos e intermediando todo o processo de acesso aos componentes físicos do computador (hardware). Outro conceito semelhante dado por Kitajima nos diz que um sistema operacional nada mais é do que um conjunto de programas (software) que executa imediatamente sobre o hardware, ou seja, gerencia o hardware e proporciona um modelo de uso simples, virtual do hardware complexo.

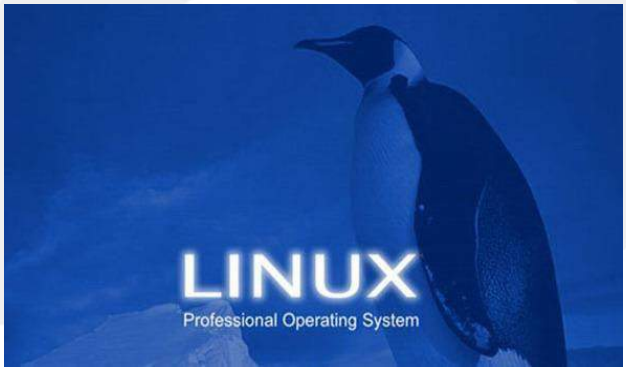
Um dado importante é que sempre que falamos de sistema operacional falamos de processos, ou seja, execução de um programa, falamos de memória, a qual guarda o estado de um processo, e falamos em gerenciadores de entrada e saída, que são os drivers.

De forma geral, um sistema operacional permite que um usuário use a CPU por meio dos processos, a RAM (memória) pelos mecanismos de gerência da memória, e use o disco, monitor, teclado, entre outros, por meio de gerenciadores de entrada e saída. Além disso, a rede também pode ser considerada um dispositivo de entrada e saída, visto que mandamos e recebemos mensagens.

Os sistemas operacionais mais conhecidos são os baseados no Windows, Unix e MacOS. O Windows é um sistema fechado, fácil de usar. Por outro lado, o Linux, variante do Unix, é um sistema aberto, que atende bem a função de gerenciar recursos físicos, como memória, CPU, entre outros. Os erros comumente vistos no Windows raramente ocorrem no Linux, porém esse último sistema apresenta

dificuldades em prover um desenvolvimento de aplicativos básicos, tais como editores de texto que sejam fáceis de usar. Em relação à bioinformática, grande parte das aplicações é compilada e distribuída para a execução em plataformas derivadas do sistema Unix, pois de forma geral, esses sistemas são mais confiáveis, e com uma capacidade de gerenciamento de grandes quantidades de dados superior ao Windows. A figura apresenta o famoso logo do sistema Linux.

FIGURA: SISTEMA OPERACIONAL LINUX E SEU FAMOSO LOGO



http://www.suggestsoft.com

É fundamental que um bioinformata ao instalar e usar um sistema operacional tenha conhecimento sobre o mesmo. Dados como especificações do hardware, quais sistemas operacionais suportam quais softwares de bioinformática e as abstrações que o sistema operacional suporta (processos), é fundamental para o bioinformata trabalhar plenamente. Um bioinformata não deve somente saber utilizar programas produzidos por outros programadores, mas também deve ser capaz de desenvolver aplicativos para lidar com os mais diversos problemas que possam ocorrer durante as análises de dados. Para isso, é de extrema importância que o profissional da bioinformática tenha alguns conhecimentos sobre linguagem de programação.

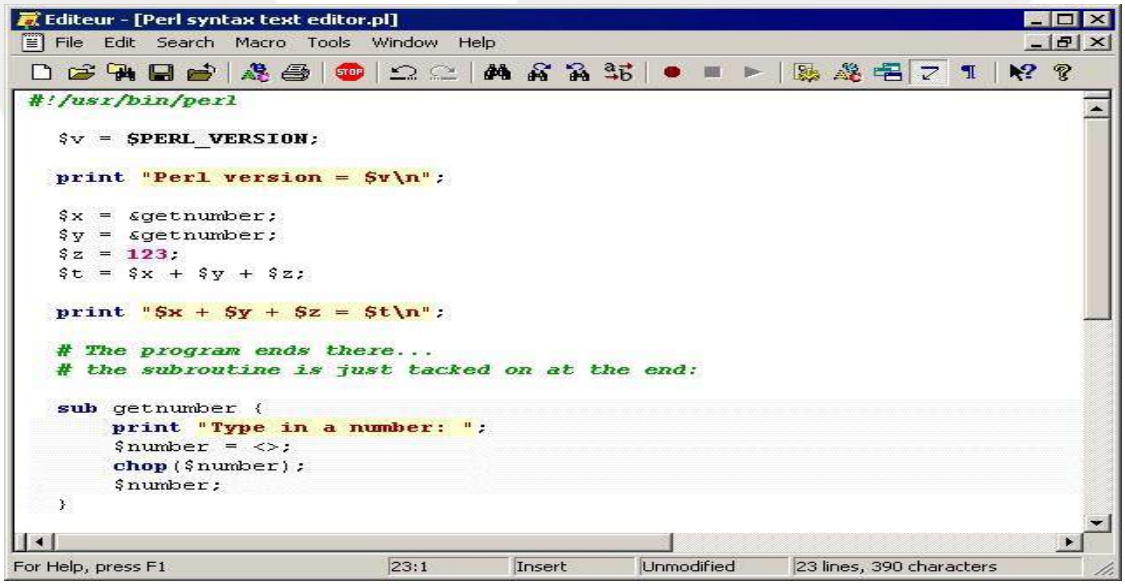
A linguagem de programação visa facilitar a especificação de tarefas a um computador. Atualmente, existem milhares de linguagens de programação utilizadas para os mais diversos fins. Essas linguagens possuem conjuntos de comandos específicos para a interface homem- computador. Dentre essas linguagens, as mais utilizadas são a basic, pascal, C, C++, Java, cobol, e fortran. No entanto, como

estamos falando de linguagens de programação em bioinformática, a mais utilizada nesse caso é a linguagem PERL (Practical Extract and Report Language).

O PERL é uma linguagem de programação simples, rica, e gratuita. O seu idealizador, Larry Wall, havia criado para produzir relatórios de informações de erros. Ao longo dos anos, porém, essa linguagem foi conquistando os usuários, sendo atualmente considerada uma linguagem sofisticada, que tem como característica forte a facilidade na manipulação de textos, e é justamente essa característica que torna o PERL a linguagem de programação mais utilizada no tratamento de dados de sequências de DNA e proteínas. Outra característica importante do PERL é fornecer uma interface de integração com bancos de dados, outro ponto que reforça seu papel como linguagem líder no seguimento de bioinformática.

Nesse sentido, o PERL apresenta uma série gratuitamente distribuídos capazes de reforçar sua funcionalidade. O PERL-DBI, por exemplo, é um conjunto que permite essa conexão com os bancos de dados mencionados anteriormente. Outros como BioPerl e Bio Graphics também apresentam ferramentas úteis para o trabalho dos bioinformatas. A figura mostra uma janela de trabalho com a linguagem de programação PERL.

FIGURA: EXEMPLO DE UMA JANELA USANDO A LINGUAGEM DE PROGRAMAÇÃO PERL



<http://guerreros-de-zion.wikispaces.com/Historia+de+los+lenguajes+de+programacion>.

BANCOS DE DADOS

O crescimento no número de dados biológicos gerados nas últimas décadas levou a uma grande mudança no cenário científico mundial, especialmente no que se refere ao uso efetivo das informações geradas. A catalogação correta, bem como a marcação e a conexão com as informações sequenciais, estruturais e funcionais de genes e proteínas facilitam a descoberta de novos circuitos biológicos e têm um papel fundamental no entendimento dos complexos sistemas de vida na terra. As informações precisam ser estocadas de forma correta e completa, de modo a facilitar a conexão com outros elementos, e acima de tudo devem ter uma interface fácil de navegar.

Ferramentas computacionais e bancos de dados são essenciais no manejo e identificação dos elementos dos bancos de dados que refletem a organização dos sistemas biológicos. O Centro Internacional de Informações Biotecnológicas (NCBI) nos Estados Unidos, e o Instituto de Bioinformática Europeu (EBI) na Inglaterra são os dois principais centros responsáveis pelo manejo dos dados biológicos a nível mundial. Ambos os centros mantêm bancos de dados confiáveis e programas analíticos que servem como valiosas ferramentas para a comunidade científica. Novos dados são adicionados diariamente e alocados apropriadamente dentro desses bancos, permitindo um intercâmbio de informações entre os pesquisadores.

As figuras mostram as páginas iniciais do NCBI e EBI, respectivamente.

FIGURA: PÁGINA INICIAL DO NCBI

<http://www.ncbi.nlm.nih.gov>.

FIGURA: PÁGINA INICIAL DO EBI

FONTE: Disponível em: <http://www.ebi.ac.uk/>..

Os serviços oferecidos pelo NCBI e EBI são possíveis graças a computadores elaborados e rápidos, os quais permitem sofisticadas análises, e à internet que facilita a comunicação eletrônica. Exemplos de uso dessas ferramentas nas ciências da vida incluem análise de sequências, desenho de pequenas moléculas para o desenvolvimento de drogas, predição da função proteica, e simulação do papel dos genes nos diversos mecanismos celulares e fisiopatogênicos.

A utilidade do NCBI e EBI, bem como de outros bancos de dados disponíveis, é atuar como um “coletor” e integrador dos mais diversos conhecimentos sobre a estrutura e função de genes e proteínas de interesse biológico e médico. Esses bancos atuam como bibliotecas eletrônicas que podem ser lidas com o uso de ferramentas de bioinformática adequadas, as quais são indispensáveis no campo científico. De modo geral, a organização dos bancos de dados de informações biológicas inicia com os dados brutos; sequências de DNA, estrutura de proteínas, e dados funcionais. Essas bases funcionais devem proporcionar uma submissão bastante acurada, bem como uma recuperação dos dados brutos, permitindo também a construção de bases de dados secundários e terciários incluindo bibliografia, anotações médicas, e interatividade para encontrar informações biológicas relevantes, tais como programas para alinhamento de sequências, identificação de genes, comparação de genomas, tradução das sequências de DNA em aminoácidos e vice-versa, identificação de homólogos e parálogos, entre outras funções.

Filho et al. afirmam que grande parte dos bancos de dados são atrelados a um sistema de gerenciamento de bancos de dados (SGBD), o qual é responsável por intermediar os processos de construção, manipulação e administração dos bancos de dados. Atualmente, existem diversos sistemas de gerenciamento de dados. Por exemplo, o mysqlé, o qual é um sistema bastante utilizado pela comunidade científica e em projetos genoma, devido ao fato de ser um sistema gratuito, de código aberto, e veloz. O postgreSQL também é um sistema gratuito, mas apresenta grandes dificuldades no seu gerenciamento, o que o torna menos utilizado. O Oraclee e o SQLServer são sistemas robustos e sofisticados, mas de alto custo, sendo dessa forma limitados às grandes empresas.

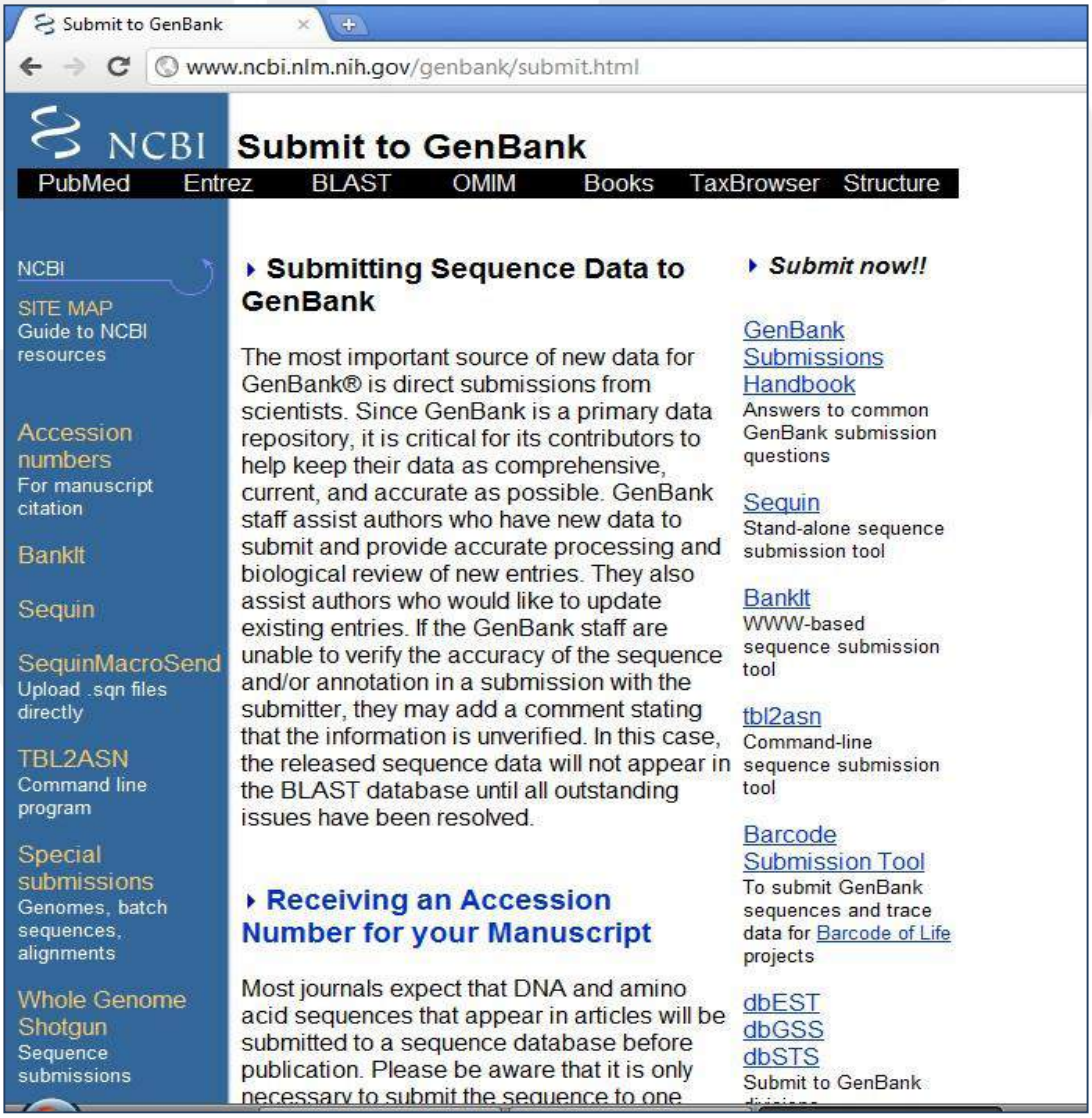
SUBMISSÃO DE DADOS

A principal fonte de informações dos grandes bancos de dados centrais são as contribuições diretas dos pesquisadores. Antigamente, não era incomum encontrar pesquisadores trocando informações sobre sequências de DNA ou aminoácidos com outros pesquisadores por meio do telefone, o que levava obviamente a um número superestimado de erros nos processos de replicação e transcrição do DNA. Atualmente, o envio de informações gênicas e proteicas é feito por e-mail e transferido para bancos como Genbank ou SWISS-PROT de forma fácil, rápida, e virtualmente isenta de erros. O desenvolvimento das ferramentas da internet facilitou e muito esse processo.

Os programas BanKIt ou o Sequin são fornecidos pelo NCBI para enviar sequências e anotações biológicas para os cientistas do Genbank, que irão criar um número de acesso para esses dados, permitindo a liberação quase que imediata para os domínios públicos (cerca de 48 horas). Na divisão de sequências genômicas do NCBI, as sequências submetidas são classificadas em três fases: não completadas e não ordenadas, não completadas e ordenadas, e sequências finais que não contêm erros sequenciais. Devido ao acelerado processo de sequenciamento e submissão de novas sequências, a identificação precoce de qualquer erro é fundamental. O NCBI, bem como outros bancos de dados públicos dispõe de mecanismos computacionais, bem como de cientistas treinados que visam detectar e corrigir esses erros.

É importante ressaltar que normalmente os pesquisadores que submetem uma sequência gênica apresentam uma sequência única desse gene e algumas informações biológicas relevantes sobre o mesmo. No entanto, quando estamos lidando com projetos genômicos, a submissão de sequências originadas de ESTs (expressed sequence tags), STSs (sequence tagged sites), e GSSs (genoma survey sequences) requerem procedimentos mais especializados, visto que essas sequências diferem das sequências gênicas e proteicas funcionais, por seu tamanho reduzido e a grande frequência que são encontradas no genoma.

Dessa forma, procedimentos adicionais e mais complexos são requeridos para garantir a reprodutibilidade das informações submetidas. Ainda, é fundamental destacar que existe um intercâmbio de informações entre os principais bancos de dados, os quais asseguram que as informações submetidas pelos pesquisadores não sejam redundantes, ou seja, que não sejam submetidas mais de uma vez. Outro dado importante sobre as ferramentas de submissão de dados, é que os autores podem atualizar as informações submetidas. A figura mostra a página inicial do NCBI para submissão de sequências.



<http://www.ncbi.nlm.nih.gov/genbank/submit.html>

ANOTAÇÃO DE DADOS E INTERCÂMBIO DE INFORMAÇÕES ENTRE OS BANCOS DE DADOS

Comumente, os cientistas costumam trabalhar em pequenos grupos independentes, o que muitas vezes resulta na identificação repetitiva de genes e proteínas. Isso é claramente observado na submissão de dados, tais como sequências de DNA, as quais são submetidas mais de uma vez e muitas vezes com nomes diferentes e anotações (dados que permitem a identificação correta do elemento de interesse) inconsistentes.

Nesses casos, apenas especialistas da área são capazes de reconhecer que essas múltiplas entradas se referem a apenas um elemento, por exemplo, um gene. Diante dessa situação, é fundamental que seja desenvolvida uma terminologia clara entre as diferentes faculdades biológicas, a fim de fornecer uma anotação padronizada e adequada de sequências e estruturas gênicas e proteicas nos bancos de dados, o que facilitará a conectividade entre os bancos e aumentará e melhorará a qualidade dos conhecimentos biológicos. Além disso, a identificação de sequências redundantes é fundamental para o controle de qualidade desses bancos.

NOTAÇÕES DE DADOS

Para entender um gene ou sua sequência, precisamos inicialmente entender o contexto, ou seja, quais as informações biológicas associadas que define a sua função, dados sobre a sua estrutura, localização cromossômica, bem como a estrutura e função do seu produto, proteína ou RNA, e sua classificação taxonômica. Um exemplo de anotações mínimas necessárias que devem acompanhar uma sequência de DNA é demonstrado na lista abaixo:

a) Sequências relacionadas nos bancos de dados;

b) Predição de estrutura/comparação com estruturas de raios-X;

c) ORF (open reading frame – estrutura de leitura aberta) se a função não é conhecida;

d) Estrutura do domínio;

e) Segmentos transmembrana;

f) Sequência sinal;

g) Sítios consenso para glicosilação, fosforilação, entre outros;

h) Nomenclatura alternativa;

i) Informações genéticas de sequências regulatórias;

j) Informações sobre o processo de tradução;

k) Peso molecular;

l) Bibliografia.

A lista acima apresenta elementos intrínsecos do registro de um gene que resulta da aplicação de ferramentas de bioinformática. Outras anotações podem ser obtidas de informações contextuais e experimentais. Uma anotação completa deve incluir todas as informações biológicas disponíveis associadas com o gene de interesse. A fim de tornar essas informações disponíveis e acessíveis, é fundamental que um vocabulário controlado (ontologia) compartilhado entre as diversas disciplinas biológicas exista.

A classificação sistemática de organismos tem tradição nas ciências da vida. No entanto, o desenvolvimento de projetos genômicos trouxe à tona a necessidade de desenvolvimento de uma linguagem comum para descrição de genes e proteínas, células e organismos. A classificação taxonômica de organismos como espécie, gênero, família, ordem, classe, filo, reino e domínio, é bem conhecida na biologia. No entanto, quando falamos de genes e proteínas, cada grupo costuma usar uma terminologia própria. Dessa forma, um novo projeto de ontologia biológica voltada para a classificação de genes e proteínas têm se desenvolvido nos últimos anos.

É notável que a informação de um único gene possa ser encontrada em vários bancos de dados, com vários nomes e anotações diferentes, o que a primeira vista pode parecer se tratar de diferentes genes e produtos gênicos, mas um olhar mais atento percebe que se trata da mesma coisa. Diante disso, recentemente a

comunidade biológica começou a usar o vocabulário criado pelo Consórcio De Ontologia Gênica (Gene Ontology Consortium) referida como ontologia gênica. Esse consórcio agrupou as proteínas dentro de três redes estruturais: processos biológicos, componentes celulares e funções moleculares. A tabela 2 mostra os principais projetos de ontologia biológica disponíveis na rede, os quais visam padronizar o vocabulário das informações moleculares.

TABELA: PRINCIPAIS PROJETOS DE ONTOLOGIA BIOLÓGICA ABERTOS

Projetos	Vocabulário Controlado	Site
Gene Ontology Consortium	Fornece três redes estruturais para descrever os produtos gênicos: processos biológicos, componentes celulares e função molecular.	www.geneontology.org
Sequence Ontology	Conjunto de termos usados para descrever características de um nucleotídeo ou sequência proteica.	song.sourceforge.net
Generic Model Organism Databases	Visualização do genoma e ferramentas de edição, ferramentas de busca literária, robusto banco de dados, ferramentas de ontologia biológica, e um conjunto de procedimentos operacionais padrão.	gmod.sourceforge.net
Standards and Ontologies for Functional Genomics	Critérios e ontologia com ênfase na descrição de experimentos genômicos funcionais.	www.sofg.org
The Microarray Gene Expression Data	Ontologia para descrição de amostras usadas nos experimentos de *microarray*.	mged.sourceforge.net
Plant Ontology Consortium	Vocabulário para estrutura de plantas, e estágios de crescimento e desenvolvimento.	www.plantontology.org

REDUNDÂNCIA NOS BANCOS DE DADOS

A redundância das informações em bancos de dados serve como controle de qualidade dos mesmos. Algumas vezes, pesquisadores de laboratórios competidores publicam sequências do mesmo gene, porém que diferem em alguns pares de bases. A questão nesse caso é se estamos diante de um mutante verdadeiro ou apenas um erro de sequenciamento? De forma geral, se o gene foi mapeado no mesmo organismo, com as mesmas vias gênicas, é provável que erros usuais técnicos expliquem as diferenças encontradas. Dessa forma, percebemos que a redundância não apenas provém do fato de que diferentes pesquisadores estão interessados na mesma proteína e gene, mas também do fato de que diferentes técnicas de clonagem e sequenciamento randômico do genoma criam fragmentos com poucas anotações biológicas relevantes.

A fim de fornecer um banco de dados consistente e sem redundância, o RefSeq do NCBI contém um conjunto não redundante de sequências do Genbank incluindo sequências de DNA genômico, RNAm, e proteínas de genes conhecidos, RNAm e proteínas de modelos genéticos e cromossomos.

Diante de tudo o que foi exposto, percebemos que com a lista crescente de genes estocados e anotados de vários organismos nos bancos de dados, é fundamental que haja interatividade entre os mesmos, a fim de fornecer informações confiáveis aos pesquisadores. No NCBI, por exemplo, foi implantada uma visão orientada do gene de interesse, que permite ao pesquisador fazer uma busca em vários bancos de dados, a não apenas no NCBI, sobre as informações relacionadas aos genes de interesse. O Entrez do NCBI funciona como um integrador, que permite a busca por meio de textos nos mais variados ambientes dessa plataforma, tais como na literatura (pubmed), nas sequências de nucleotídeos e proteínas, na estrutura proteica, genoma, taxonomia, e bancos de dados de expressão.

BANCOS DE DADOS PÚBLICOS

Os grandes investimentos na construção de bancos de dados públicos são uma das razões do sucesso dos projetos genoma, particularmente do projeto genoma humano. Diante disso, a organização sistemática de todos esses dados em grandes bancos disponíveis para acesso on-line é fundamental.

Os bancos de dados podem ser classificados basicamente em bancos primários e secundários. Os bancos contendo sequência de nucleotídeos, aminoácidos e estrutura de proteínas são denominados de primários, formados pela deposição direta de sequências, sem qualquer processamento ou análise. O Genbank do NCBI, assim Omo o EBI, o DDBJ (DNA Data Bank of Japan) e o PDB (Protein Data Bank) são os principais bancos de dados primários. Por outro lado, os bancos secundários são àqueles derivados dos primários, ou seja, formaram-se usando as informações depositadas nos bancos primários. Exemplos de bancos secundários incluem o PIR (Protein Information Resource), o SWISS-PROT, entre outros.

Os bancos podem ainda ser classificados em estruturais e funcionais. Os bancos estruturais mantêm dados relacionados à estrutura das proteínas propriamente ditas. Os funcionais por outro lado, tal como o KEGG (Kyoto Encyclopedia of Genes and Genomes), disponibiliza links para mapas metabólicos de organismos com genoma completamente ou parcialmente completos. Na tabela 3 estão listados os principais bancos de dados utilizados em bioinformática, dos quais daremos mais ênfase ao NCBI e EBI.

TABELA: PRINCIPAIS BANCOS DE DADOS UTILIZADOS EM BIOINFORMÁTICA

Bancos de Dados	Descrições	Sites
Genbank (NCBI)	Banco de dados de sequências de DNA e proteínas americano.	www.ncbi.nlm.nih.gov
EBI	Banco de dados de sequências de DNA europeu.	www.ebi.ac.uk

DDBJ	Banco de dados de sequências de DNA japonês.	www.ddbj.nig.ac.jp
PDB	Armazena estruturas tridimensionais de proteínas.	www.rcsb.org/pdb
TIGR Databases	Bancos com informações de genomas de vários organismos.	www.tigr.org/tdb
PIR	Bancos de proteínas.	www-nbrf.georgetown.edu
SWISS-PROTs	Armazena sequências de proteínas e suas características moleculares.	www.expasy.ch/spro
INTERPRO	Bancos de dados de famílias, domínios e assinaturas de proteínas.	www.ebi.ac.uk/interpro
KEGG	Bancos com sequências de genomas de vários organismos e informações relacionadas às suas vias metabólicas.	www.genome.ad.jp/kegg

NCBI

Em 1988, o senado americano reconheceu a necessidade de se ter um processamento de dados computadorizado para conduzir as pesquisas biomédicas, e patrocinou o projeto legislativo que ajudou a estabelecer o Centro Nacional de Informações Biotecnológicas (NCBI) como uma divisão da Biblioteca Nacional de Medicina (NLM) do Instituto Nacional de Saúde (NIH). O foco do NLM é a manutenção das bases de dados biomédicas, enquanto que o NCBI está diretamente envolvido no desenvolvimento de novas ferramentas analíticas que permitam um melhor entendimento dos processos genéticos e moleculares relacionados aos principais eventos patogênicos. De forma geral, os quatro principais objetivos do NCBI são: criar máquinas automatizadas que possam analisar e estocar permanentemente dados de biologia molecular, genética e bioquímica, facilitar o uso dos bancos de dados e dos programas analíticos pela comunidade científica, coordenar esforços mundiais para reunir dados biológicos, e

conduzir pesquisas computacionais da relação estrutura-função das principais moléculas biológicas.

Para conseguir cumprir seus objetivos primordiais, o NCBI é composto de profissionais das mais diversas áreas do conhecimento, tais como computação, biologia molecular, matemática, bioquímica, física e biologia estrutural. A colaboração de diversos profissionais do NCBI permite que sejam realizados estudos das bases moleculares das doenças por meio de ferramentas matemáticas e computacionais. Sendo assim, as principais áreas desses estudos incluem a análise de sequências gênicas ou de produtos gênicos de interesse, o melhor entendimento da organização estrutural gênica, bem como a predição estrutural das moléculas analisadas, por exemplo, proteínas.

O NCBI trabalha com vários bancos de dados, dos quais se destacam os bancos de dados de sequências de proteínas, os quais incluem bancos de proteínas redundantes como PIR, e não redundantes como NR, SWISS-PROT e PDB; além dos bancos de dados de sequências de nucleotídeos, tanto DNA como RNA, os quais também podem ser redundantes como o dbEST, ou não redundantes como o GenBank. A questão da redundância em base de dados biológicos é algo bastante complicado como explicado no capítulo anterior. De forma geral, cada base de dados tem seus próprios critérios para estabelecer o que considera redundância ou não. Muitas dessas bases utilizam métodos automáticos para avaliar esse aspecto, o que de certa forma é menos sensível que a avaliação manual, porém permite que um número maior de dados seja analisado em um tempo menor.

Os principais bancos de dados de sequências de proteínas do NCBI são: Alu (conjunto de repetições Alu traduzidas), E.coli (base de dados específica para o genoma codificador de E.coli), Kabat (base de dados voltada para sequências de proteínas com interesse imunológico),

Month (base de dados contendo as mais recentes sequências gênicas traduzidas do GenBank, PDB, SWISS-PROT e PIR), NR (base de dados de todas as sequências traduzidas não redundantes do GenBank, PDB, SWISS-PROT e PIR), PDB (base de dados de proteínas, nas quais as estruturas em três dimensões são

conhecidas), e SWISS-PROT (base de dados de sequências de proteínas não redundante, considerada uma das mais informativas disponíveis).

Por outro lado, os principais bancos de sequências de nucleotídeos são: Alu (base de dados de sequências nucleotídicas repetidas Alu), dbEST (base de dados não redundante questionável do GenBank, EMBL, e DDBJ EST), dbSTS (base de dados não redundantes do GenBank, EMBL, e DDBJ EST), E.coli (base de dados de sequências nucleotídicas específicas para o genoma de E.coli), EPD (base de dados de sequências promotoras de organismos eucarióticos), GSS (base de dados que contém sequências de nucleotídeos exon-trapped, sequências de PCR Alu, entre outras), HTGS (base de dados de sequências genômicas de alta capacidade), Kabat (base de dados de sequências de nucleotídeos com interesse imunológico), Mito (base de dados de sequências mitocondriais), Month (base de dados contendo as mais recentes sequências de nucleotídeos encontradas no GenBank, EMBL, DDBJ, e PDB), NR (base de dados não redundante de sequências nucleotídicas do GenBank, EMBL, DDBJ, e PDB), e PDB (base de dados contendo sequências derivadas de estruturas tridimensionais de proteínas).

O NCBI oferece uma série de serviços à comunidade científica, tais como Pubmed (Public Medline), BLAST (Basic Local Alignment Search Tool), Entrez, BankIt (Worldwideweb submission), OMIN (Online Mendelian Inheritance in Man), entre inúmeros outros.

Resumidamente, podemos afirmar que o pubmed é um serviço de busca do NLM, que permite aos usuários o acesso a milhares de publicações biomédicas das bases de dados do MEDLINE e do preMEDLINE; o Entrez é considerado umas das ferramentas mais populares do NCBI, e permite que o usuário tenha acesso a citações bibliográficas e dados biológicos de uma grande variedade de bancos de dados, tais como PDB, SWISS-PROT, MEDLINE, entre outros; e o BankIt é o servidor de submissão de sequências ao GenBank. Ainda, o BLAST é um conjunto de programas robustos, capazes de analisar sequências de DNA e proteínas, desenhados para identificar potenciais homologias; e o OMIN é uma base de dados de genes e doenças genéticas humanas. Os estudos de taxonomia contam com uma página específica do NCBI, a qual contém bancos de dados de organismos

com nomes científicos, dos quais as sequencias são conhecidas. Por último, a home Page do NCBI “Structure” compreende a base de dados de modelagem molecular (MMDB) e uma variedade de ferramentas computacionais relevantes para a análise estrutural.

A partir de agora vamos ver na prática como funcionam as principais ferramentas do NCBI. Para isso, considere o seguinte exemplo:

a) Um aluno deseja pesquisar o termo “insulina” no NCBI. Inicialmente ele digita o endereço http://www.ncbi.nlm.nih.gov/, e na página inicial seleciona a opção “all databases” e clica em “search.

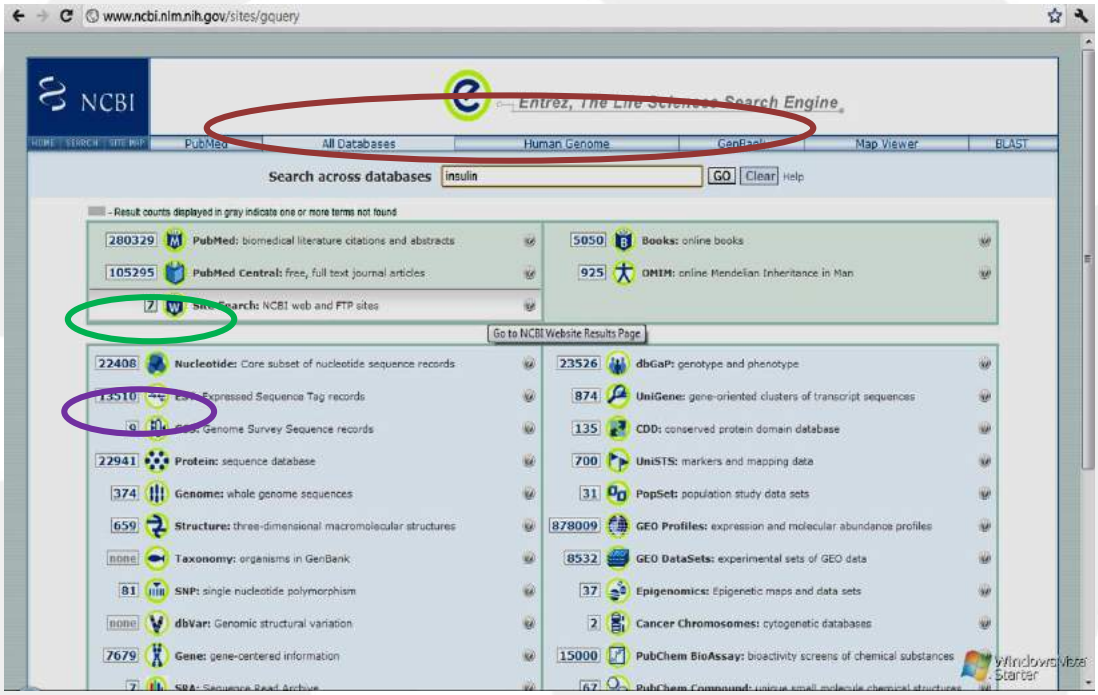
FIGURA: PÁGINA INICIAL DO NCBI. DESTAQUE PARA A OPÇÃO “ALL DATABASES” E “SEARCH”

<http://www.ncbi.nlm.nih.gov/>

b) Depois de executar o “search”, abre a página do “Entrez” que como explicado anteriormente é a integração das mais variadas bases de dados do NCBI. Na página do “Entrez”, o aluno digita o termo de pesquisa, nesse caso, “insulin” e clica em “go”. A pesquisa gerada é então realizada em uma variedade de bancos de

dados, e o número de resultados encontrados em cada base aparece à esquerda de cada recurso. Por exemplo, utilizando o termo “insulin” obtivemos 22400 resultados da base de dados de nucleotídeos, ou 22941 da base de dados de proteínas.

FIGURA: PÁGINA DO ENTREZ

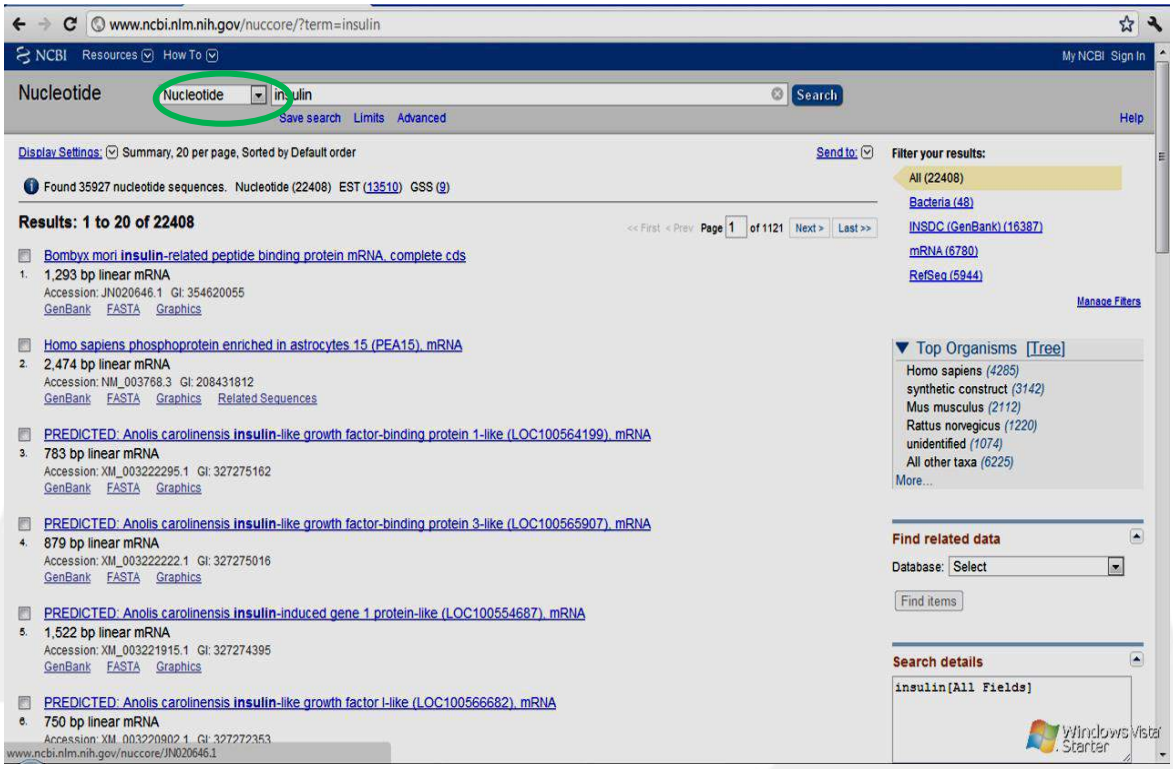


Destaque para a opção “search across databses”, na qual se deve digitar o termo de interesse; e para a marcação em verde da base de dados de nucleotídeos e em lilás para a base de dados de proteínas.

<http://www.ncbi.nlm.nih.gov/sites/gquery>

c) Agora podemos percorrer as mais diversas bases de dados procurando informações sobre o assunto de interesse. Por exemplo, se clicarmos na base de dados de nucleotídeos, outra janela se abrirá com as mais diversas sequências nucleotídicas desse peptídeo.

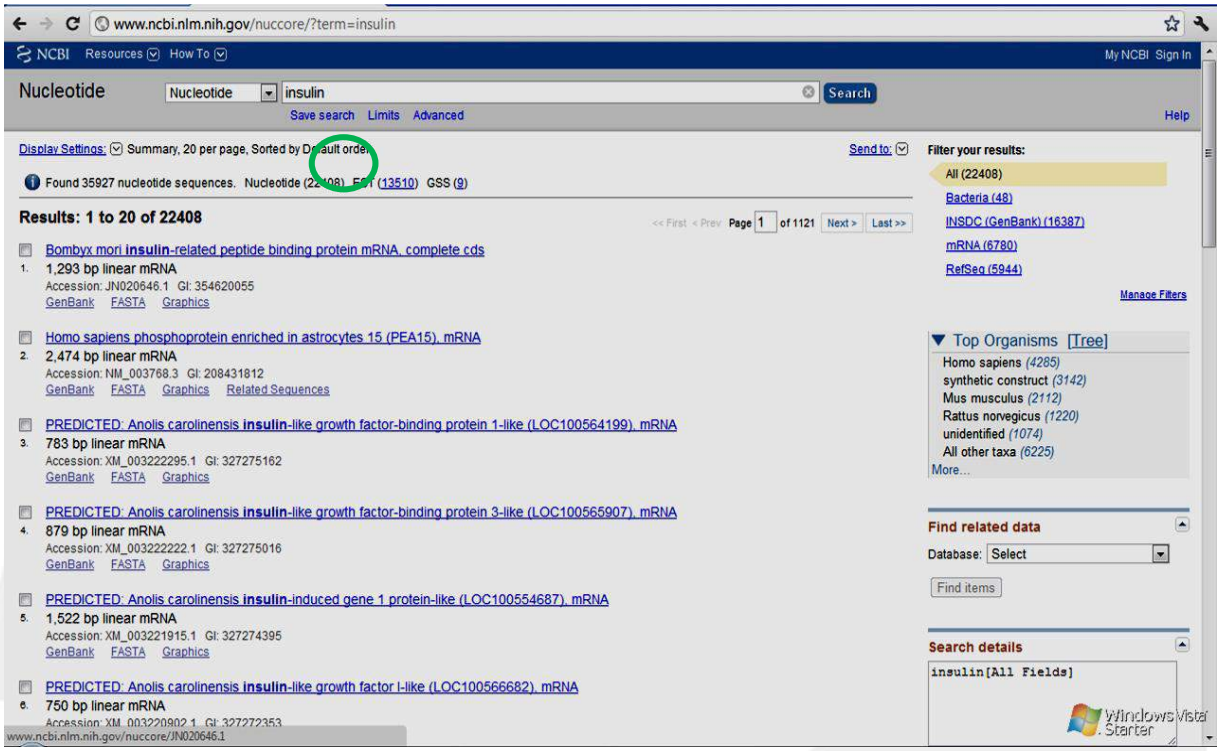
FIGURA: PÁGINA DA BASE DE DADOS DE NUCLEOTÍDEOS RELACIONADOS AO TERMO “INSULIN” DIGITADO ANTERIORMENTE.



<http://www.ncbi.nlm.nih.gov/nuccore/?term=insulin>.

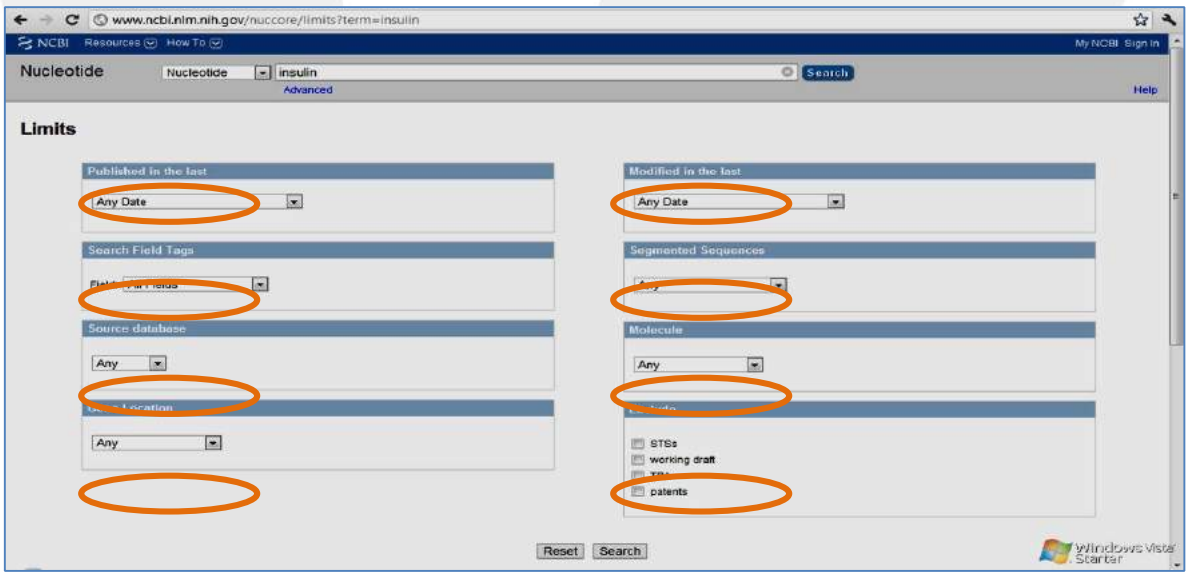
d) Como o número de resultados gerados é muito grande, podemos restringir a nossa busca por meio de algumas ferramentas disponíveis na própria página. Por exemplo, no alto da página, há uma opção denominada de “limits”. Clicando nela, abre outra janela, na qual é possível fazer restrições a nossa busca.

FIGURA: PÁGINA DA BASE DE DADOS DE NUCLEOTÍDEOS RELACIONADOS AO TERMO “INSULIN” DIGITADO ANTERIORMENTE COM DESTAQUE PARA O LINK DE RESTRIÇÃO DA PESQUISA “LIMITS”



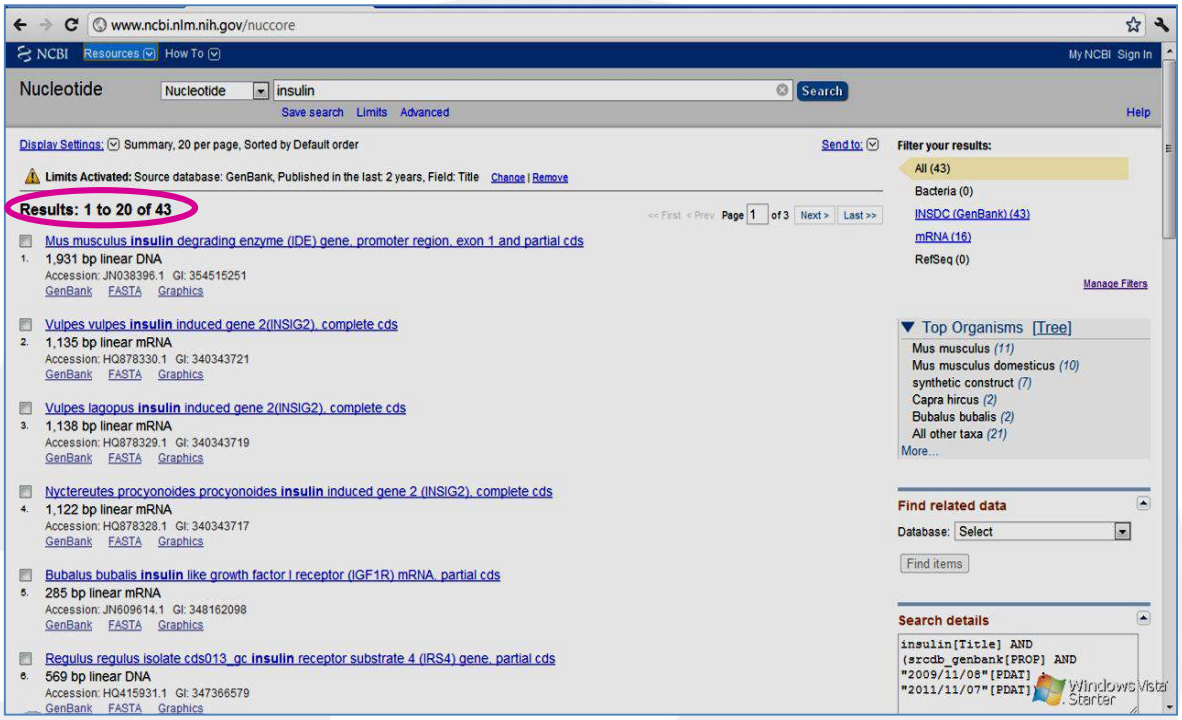
<http://www.ncbi.nlm.nih.gov/nuccore/?term=insulin>

FIGURA: PÁGINA CONTENDO AS OPÇÕES DE RESTRIÇÃO DA PESQUISA, AS QUAIS ESTÃO DESTACADAS EM LARANJA



<http://www.ncbi.nlm.nih.gov/nuccore/limits?term=insulin>.

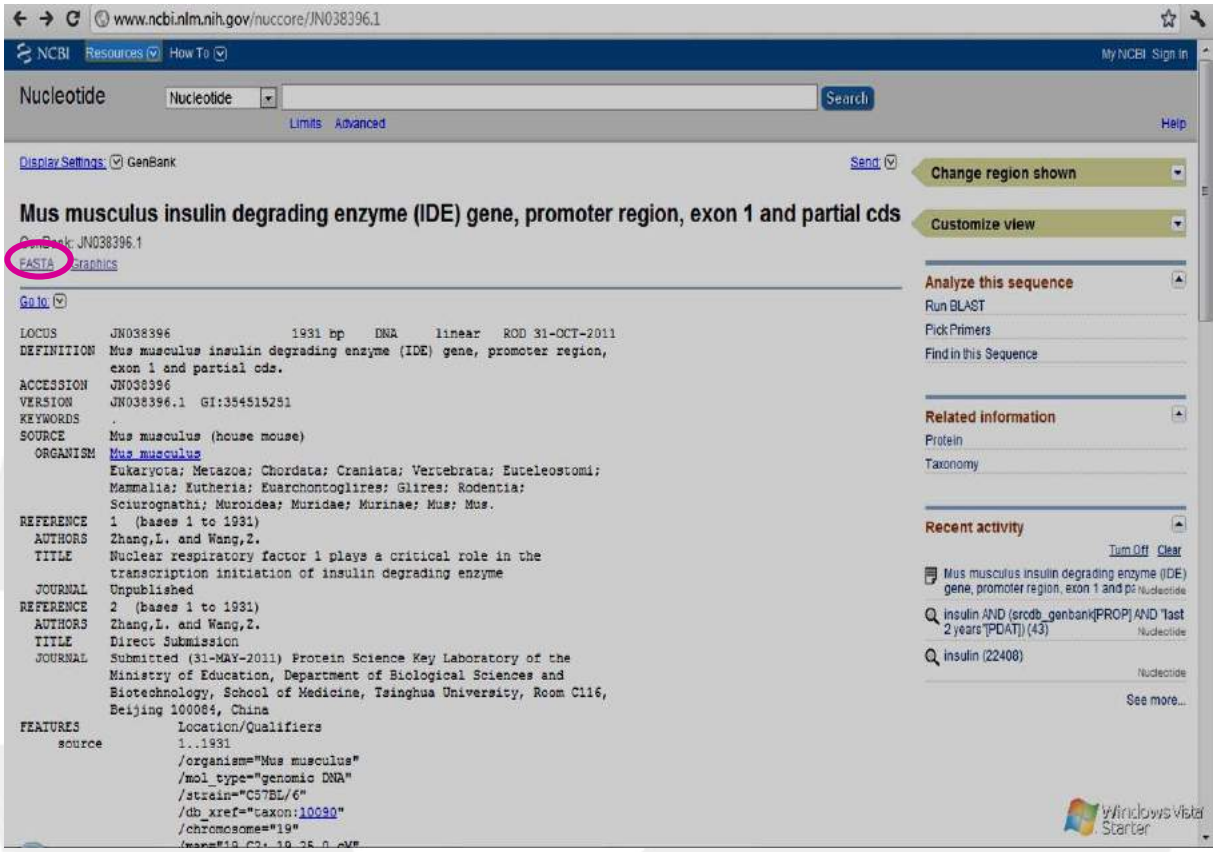
FIGURA: PÁGINA DA BASE DE DADOS DE NUCLEOTÍDEOS APENAS COM AS SEQUÊNCIAS DE INTERESSE



<http://www.ncbi.nlm.nih.gov/nuccore>.

e) Dependendo dos interesses científicos de cada pesquisador, inúmeras combinações de restrições podem ser realizadas. No caso do aluno que está pesquisando sobre insulina, podemos restringir a busca às publicações dos últimos dois anos apenas no campo “published in the last”; restringir a busca àquelas sequências, na qual a palavra “insulin” aparece no título, para isso trocaremos a opção “all Field” por “title” no campo “search Field tag”, e restringir a busca àquelas sequências disponíveis apenas na base de dados do GenBank no campo “source database”, e depois clicamos em “go”. Com a restrição de todos os dados acima, chegaremos a um total de 43 resultados contra 22400 da análise sem restrição, o que torna a avaliação muito mais fácil.

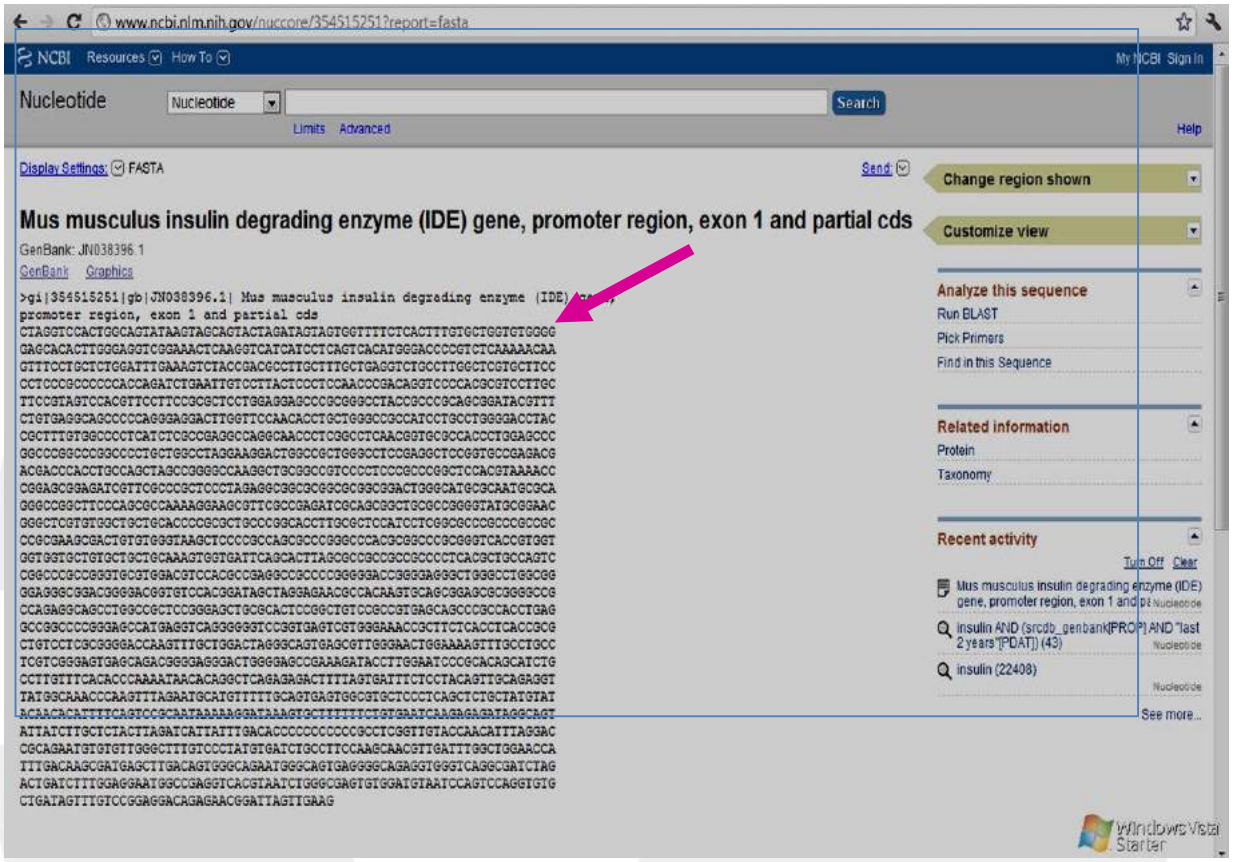
FIGURA: INFORMAÇÕES SOBRE A SEQUÊNCIA CLICADA E DESTAQUE PARA A OPÇÃO FASTA



<http://www.ncbi.nlm.nih.gov/nuccore/JN038396.1>.

f) Caso o aluno clique no primeiro item da lista, outra janela com as informações específicas daquela sequência selecionada se abrirá). Nessa página, ele pode clicar na opção fasta, e uma nova janela, agora contendo apenas a sequência no formato fasta, o qual é bastante utilizado em programas de bioinformática, aparece, facilitando com isso o trabalho do pesquisador.

FIGURA: INFORMAÇÕES SOBRE A SEQUÊNCIA CLICADA EM FORMATO FASTA



<http://www.ncbi.nlm.nih.gov/nuccore/354515251?report=fasta>.

Outros inúmeros recursos estão disponíveis no NCBI. No exemplo acima, clicamos em apenas uma base de dados e aprendemos como limitar buscas. Poderíamos ter clicado em qualquer outra base, realizando outras modificações, enfim, as opções de análise são inúmeras. O objetivo é apenas dar uma ideia de como funciona o NCBI, de forma alguma pretendemos abordar todos os tópicos disponíveis. Recomendo que acessem o site do NCBI (http://www.ncbi.nlm.nih.gov) e façam buscas diversas, clicando nas mais variadas opções e tentando entender as ferramentas disponíveis para cada uma delas. É um trabalho árduo, que exige bastante dedicação e paciência, mas posso garantir que é a única forma de aprender bioinformática, ou seja, “colocando a mão na massa”.

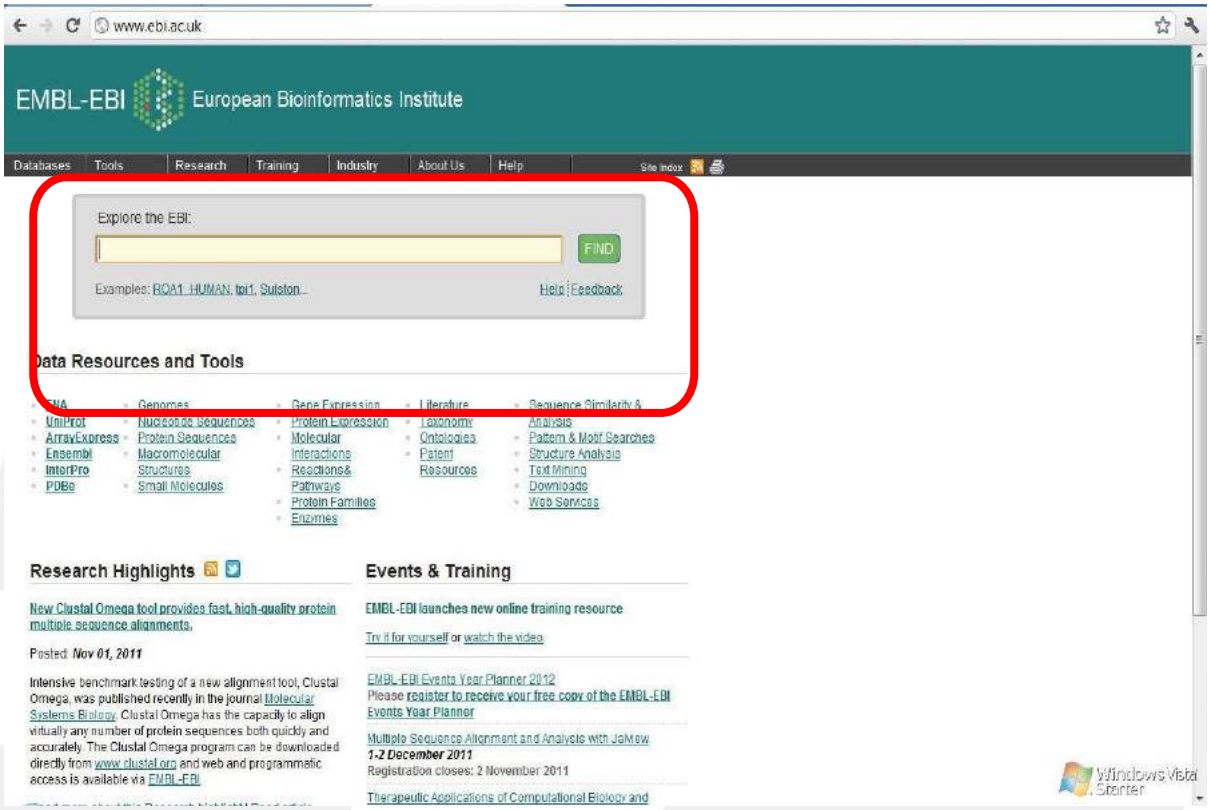
EBI

O Instituto de Bioinformática Europeu (EBI) faz parte do Laboratório de biologia Molecular Europeu (EMBL) com sede na Inglaterra. O objetivo central do EBI é conduzir pesquisas e fornecer informações sobre bioinformática para a comunidade científica mundial. O EBI é comparável ao NCBI nos Estados Unidos, e é o principal servidor de bioinformática do continente europeu.

Apresenta propósitos semelhantes ao NCBI, tais como o desenvolvimento de novas tecnologias de bioinformática, pesquisa e desenvolvimento de novos programas de bioinformática, treinamento e suporte aos assinantes, e prestar serviços relevantes de bioinformática aos pesquisadores. O EBI, assim como o NCBI, é composto de profissionais de áreas diversas, como computação, física, matemática, estatística, biologia molecular e estrutural, entre outras. Esse quadro multidisciplinar permite que sejam realizados estudos moleculares das bases das doenças, por meio de ferramentas matemáticas e computacionais.

De forma geral, as principais pesquisas desenvolvidas pelo EBI são o desenvolvimento de algoritmos de comparação mais robustos, bem como a criação de um sistema de informações mais elaborado, integrador, e fácil de usar, e o desenvolvimento de bancos de dados mais eficientes. Diante disso, os principais serviços prestados pelo EBI são as bases de dados, a submissão de dados, análise de similaridade (BLAST E FASTA), aplicativos on-line, arquivos FTP, pesquisa e desenvolvimento. As principais bases de dados do EBI são: EMBL, Uníparos/SWISS-PROT, Ensemble, MSD/PDB, IMGT, entre outras. A figura mostra a página inicial do EBI com todas as suas ferramentas.

FIGURA: PÁGINA INICIAL DO EBI COM DESTAQUE PARA AS SUAS PRINCIPAIS FERRAMENTAS



<http://www.ebi.ac.uk/>.

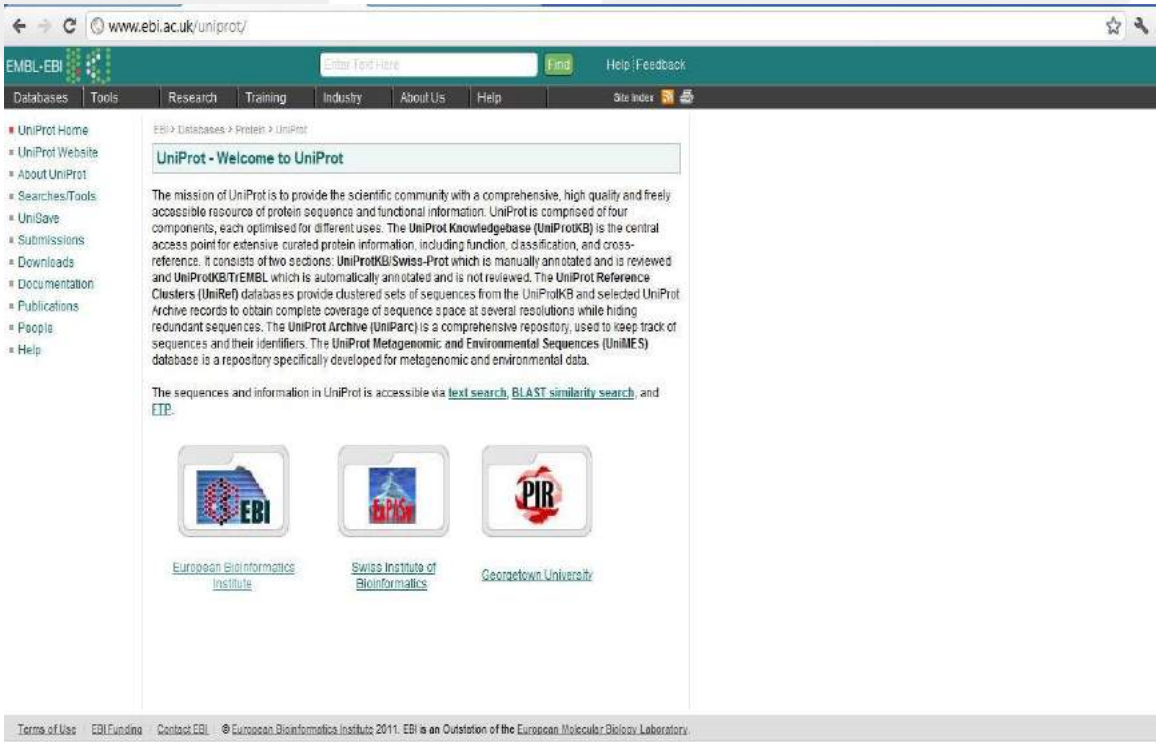
O EMBL é uma base de dados de sequências nucleotídicas (DNA ou RNA), provenientes de diversas fontes, tais como da literatura científica e de patentes, mas principalmente da submissão direta de sequências pelos pesquisadores ou grupos de sequenciamento. Essa base de dados é uma colaboração entre o GenBank do NCBI e o DDBJ do Japão. Essas três bases de dados trocam informações diariamente, mantendo o conteúdo mais atualizado possível. Resumidamente, os arquivos de sequências nucleotídicas do EMBL oferecem as seguintes informações: sequência propriamente dita; descrições teóricas dessa sequência; fonte, ou seja, a que organismo pertence essa sequência; referências bibliográficas; localização de regiões codificadoras nas sequências; e possíveis sítios de significado biológico na sequência de interesse. Vários projetos genômicos colaboram com o EBI, tais como

o projeto genoma humano (Homo sapiens), de nematoides (C.elegans), Camundongo (M.musculus), Yeast (S.pombe), entre outros.

O UniProt é uma grande base de dados de informação proteica, servindo como um repositório central de sequências de proteínas do EBI. O UniProt é composto de três partes distintas: o UniProt propriamente dito que é o ponto de acesso central para informações de proteínas, as quais incluem função, classificação e referências cruzadas; o UniRef que é uma base de dados de referências não redundantes do UniProt; e o UniParc que é um arquivo do UniProt, o qual apresenta a história de todas as sequências proteicas já depositadas no UniProt.

É interessante mencionar que a Universidade de Geneva e o EBI mantêm em conjunto a base de dados de proteínas SWISS-PROT. Sequências de DNA do EMBL são traduzidas e diretamente submetidas ao SWII-PROT, o qual é uma adaptação da base de dados PIR. O SWISS-PROT é uma base de dados não redundante e apresenta referências cruzadas com uma série de outras bibliotecas, ou seja, por meio do SWISS-PROT podem-se acessar outras bases como EMBL, PDB entre outras. A figura mostra a interface do UniProt.

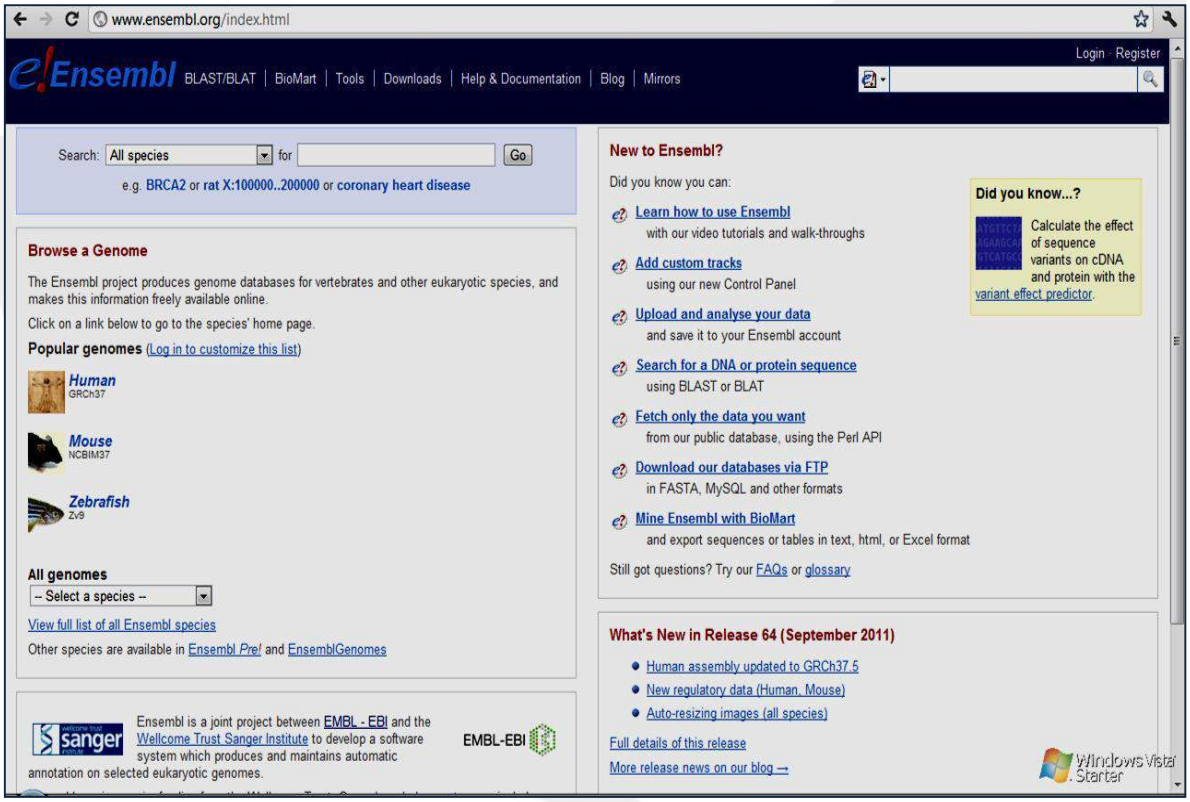
FIGURA: INTERFACE DO UNIPROT COM SUAS FERRAMENTAS OPERACIONAIS



<http://www.ebi.ac.uk/uniprot/>.

O Ensembl é um projeto do EMBL-EBI juntamente com o Instituto Sanger. O Ensembl é uma grande base de dados de genomas completos, o qual contém informações confiáveis com predições gênicas confirmadas e integradas de diversas fontes. O Ensemble contém informações de genes conhecidos, mas também prediz novos genes, por meio de anotações funcionais de famílias gênicas do InterPro (base de dados de proteínas), OMIM (base de dados sobre a herança Mendeliana em humanos) e SAGE (base de dados de análise seriada da expressão gênica). O Ensemble apresenta uma interface fácil de usar e uma série de integrações com outras bases de dados, sendo por isso bastante utilizado cientificamente.

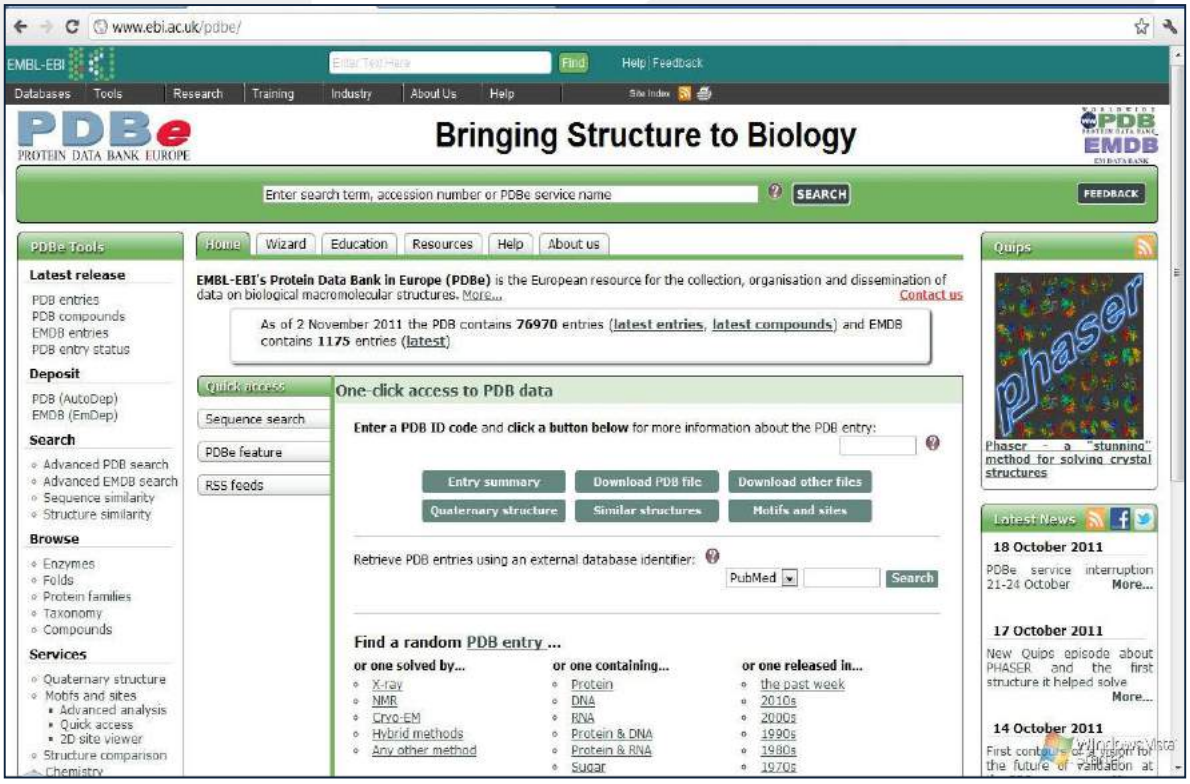
FIGURA: PÁGINA INICIAL DO ENSEMBLE



<http://www.ensembl.org/index.html>.

A base de dados de estrutura macromolecular apresenta dados de estrutura proteica do banco de dados de proteínas (PDB). O PDB é uma biblioteca de todas as estruturas tridimensionais de domínio público conhecidas. O PDB tem informações de estruturas de proteínas, proteínas mais sequências de nucleotídeos, proteínas complexadas com metais, proteínas complexadas com inibidores. As estruturas tridimensionais são determinadas basicamente de duas formas: ressonância nuclear magnética (RMN), na qual a estrutura é determinada em solução e, por isso dá mais informações sobre a suas características funcionais, e por cristalografia de raio-x, na qual a estrutura é determinada na forma cristalizada, ou seja, estática, fornecendo com isso menores informações funcionais. Nos arquivos do PDB temos informações sobre as coordenadas atômicas determinadas pela RMN ou cristalografia, citações bibliográficas, informações sobre a estrutura primária e secundária, fatores estruturais cristalográficos e dados experimentais da RMN. A figura ilustra a página inicial do PDB europeu.

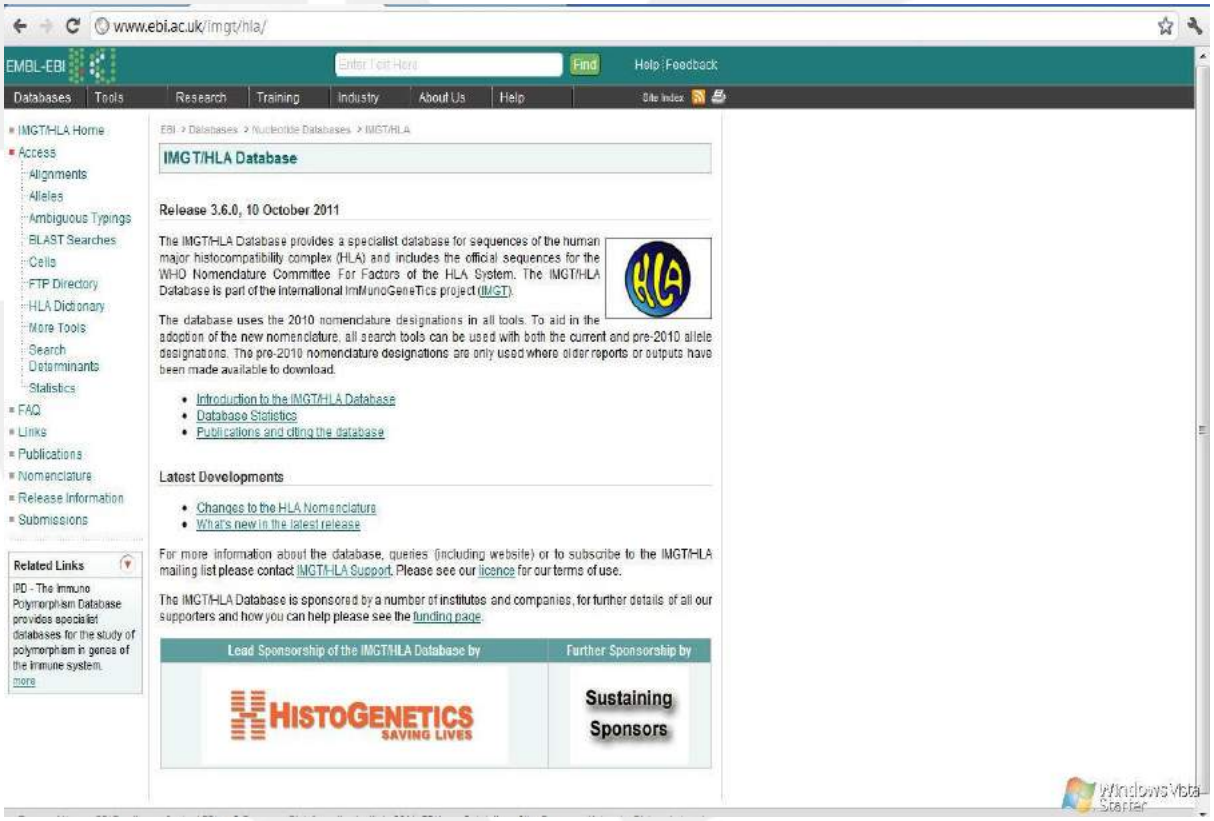
FIGURA: PÁGINA INICIAL PDB



<http://www.ebi.ac.uk/pdbe/>.

Por último, o IMGT (International ImMunoGeneTics Database) é uma base de dados voltada para genes importantes imunologicamente, muitos dos quais pertencentes à família das imunoglobulinas. Esse banco nos dá informações sobre as sequências nucleotídicas e de proteínas, alinhamentos de sequências, informações sobre alelos, polimorfismos, mapas genéticos e doenças relevantes conhecidas. A figura mostra a interface do IMGT/HLA que é umas das divisões do IMGT.

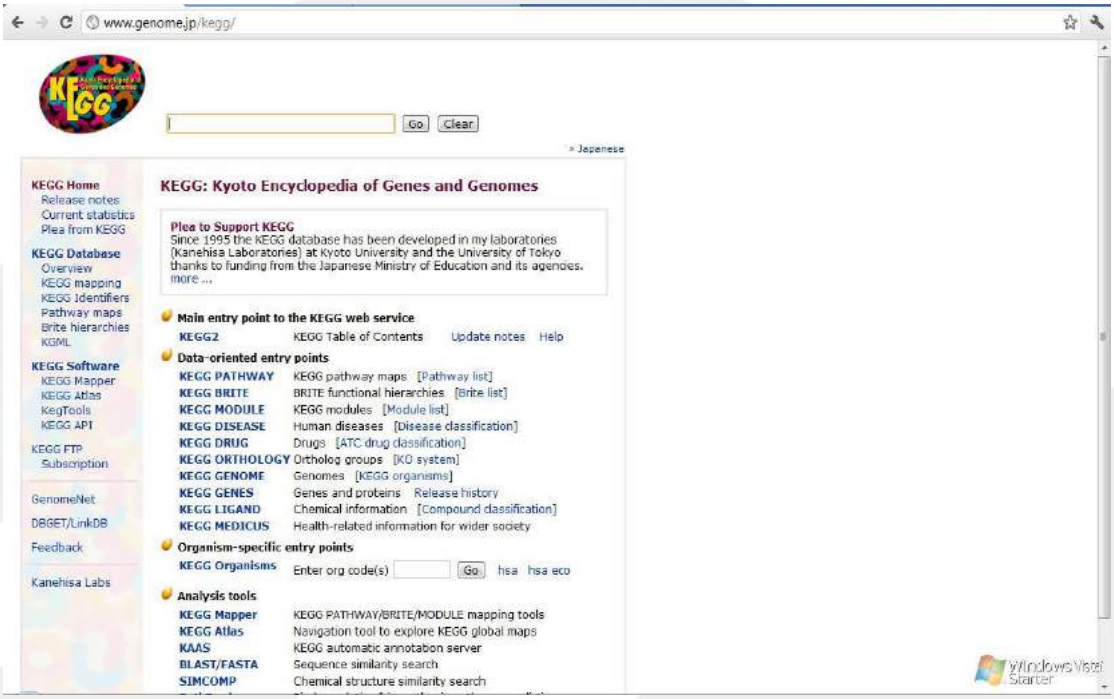
FIGURA: INTERFACE DO IMGT/HLA.



Para finalizar e a título de curiosidade, vale citar a base de dados “Kyoto Encyclopedia of Genes and Genomes” (KEEG) japonesa. O KEEG faz parte do GenomeNet que é uma rede de bases de dados e serviços computacionais para pesquisa genômica e de áreas relacionadas em biologia molecular e celular japonesa. O KEEG fornece uma base de dados metabólicos únicas e interatividade

entre as bases de dados de sequências e estruturais do GenBank e EBI. O KEEG é voltado para a identificação de rotas metabólicas em diversos organismos.

FIGURA: PÁGINA INICIAL DO KEEG



BLAST

A comparação de sequências biológicas é uma das mais importantes operações da bioinformática. Organismos recém-sequenciados são comparados com organismos já sequenciados, a fim de se determinar as suas funções e propriedades.

O alinhamento de sequências parte do seguinte pressuposto: se duas sequências de DNA são similares, a chance de elas terem funções genéticas semelhantes são maiores, e desse modo à comparação de sequências pode nos levar a fazer inferências significativas sobre as funções e relações evolutivas.

Em termos matemáticos, podemos dizer que a distância d (x, y) é o menor número de operações de inserção, remoção, substituição, capazes de transformar a sequência x na sequência y. O alinhamento global consiste em alinhar duas sequências por inteiro, na qual os alinhamentos com maiores escores são considerados os melhores. Por outro lado, o alinhamento local procura por regiões

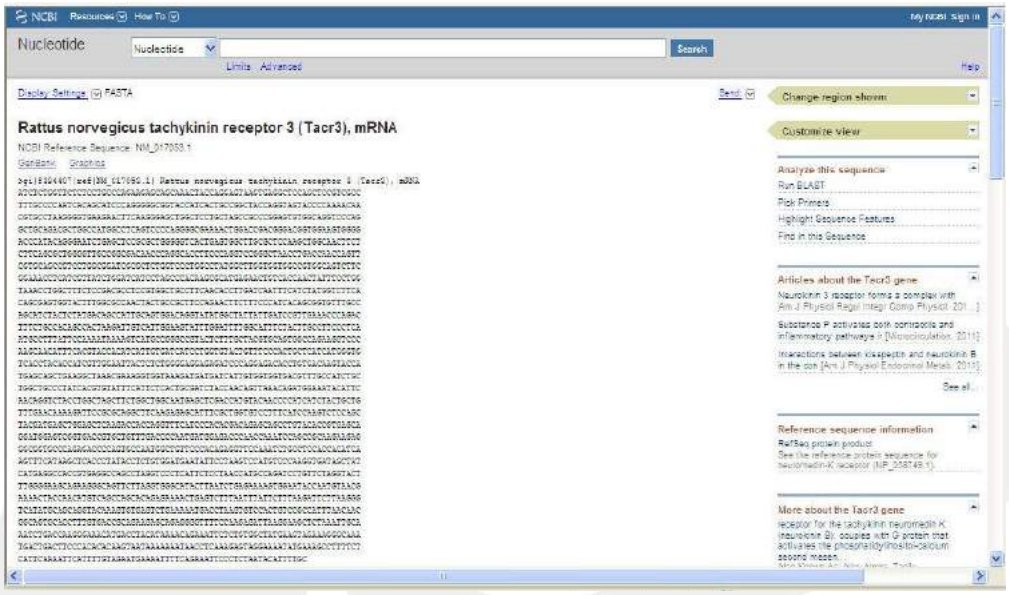
de alta similaridade dentro das sequências, mostrando com isso, regiões que possam ter se conservado ao longo da evolução, bem como regiões que possam ter funções semelhantes em genes diferentes de diferentes espécies. Diversos programas de alinhamento estão disponíveis, destacando-se o FASTA e o BLAST. Nesse curso, abordaremos o BLAST que é mais comum na rotina científica, e de interface mais fácil de usar.

Como mencionado anteriormente, o BLAST (Basic Local Alignment Search Tool) é um dos programas de alinhamento mais utilizados cientificamente. Esse programa busca regiões de similaridade entre as sequências, comparando sequências de nucleotídeos ou proteínas com sequências disponíveis nas bases de dados e calcula a significância estatística entre elas. O BLAST pode ser usado para inferir relações funcionais e evolutivas entre as sequências, assim como ajudar a identificar membros de famílias gênicas.

A seguir segue um exemplo prático de como usar o BLAST para encontrar similaridade nas sequências de nucleotídeos. No entanto, antes de iniciar o uso do BLAST, devemos ter a sequência que queremos alinhar em formato “FASTA”. Para isso, devemos seguir os passos para busca em bases de dados citadas detalhadamente. Com a sequência em formato fasta em mãos, podemos iniciar o uso do BLAST.

No exemplo dado a seguir, realizei todos os passos já descritos, ou seja, entrei no site do NCBI, depois no Entrez Pubmed coloquei meu termo de interesse “neurokinin B”, cliquei na opção “nucleotide”, e selecionei minha sequência de interesse “Rattus norvegicus tachykinin receptor 3 (Tacr3), mRNA”.

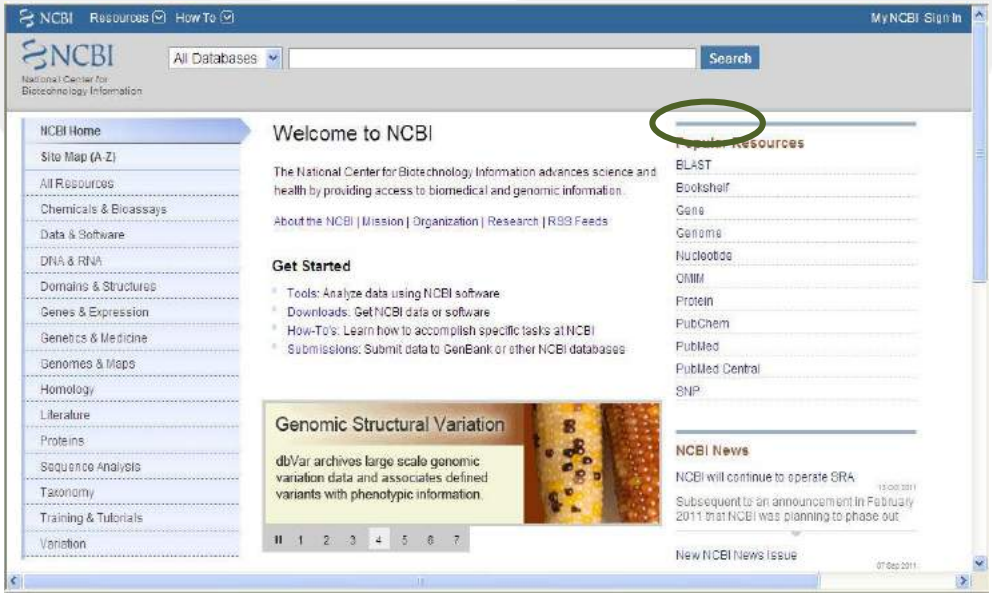
FIGURA: SEQUÊNCIA DO RATTUS NORVEGICUS TACHYKININ RECEPTOR 3 (TACR3), MRNA NO FORMATO FASTA



<http://www.ncbi.nlm.nih.gov/nuccore/8394407?report=fasta>.

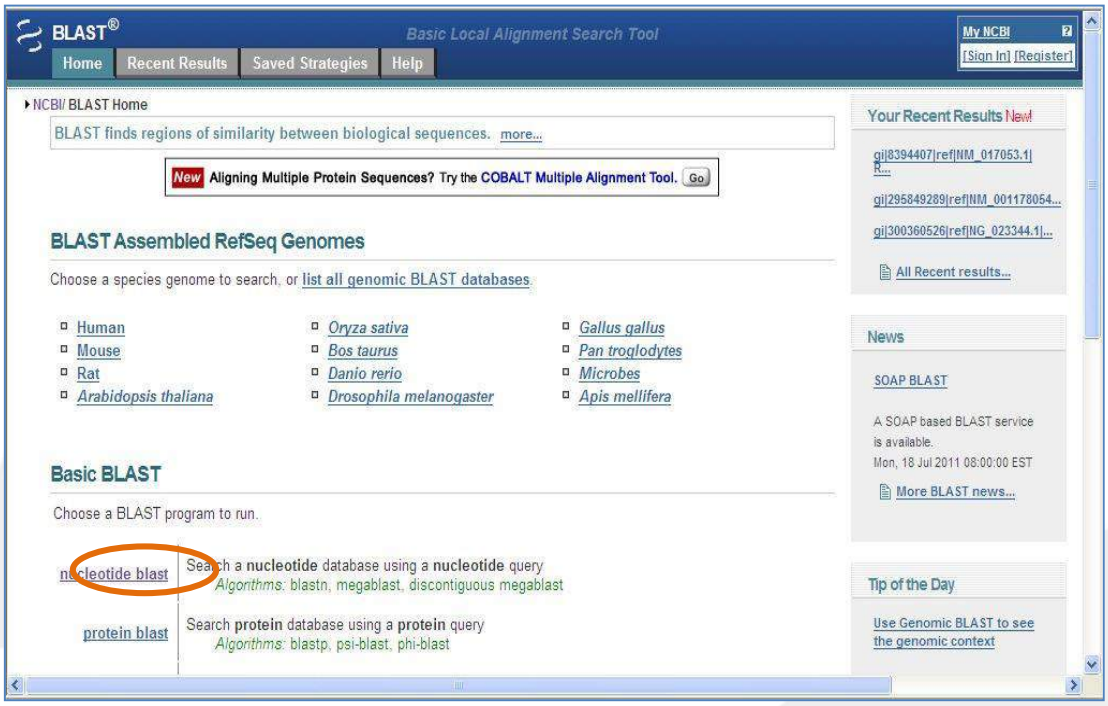
Depois de termos a sequência de interesse em formato FASTA, entraremos no site do NCBI (http://www.ncbi.nlm.nih.gov), e clicaremos no link BLAST. Na página do BLAST, clicaremos em “nucleotide blast” para buscar similaridade nas sequências de nucleotídeos.

FIGURA: PÁGINA INICIAL DO NCBI COM DESTAQUE PARA A OPÇÃO DO BLAST



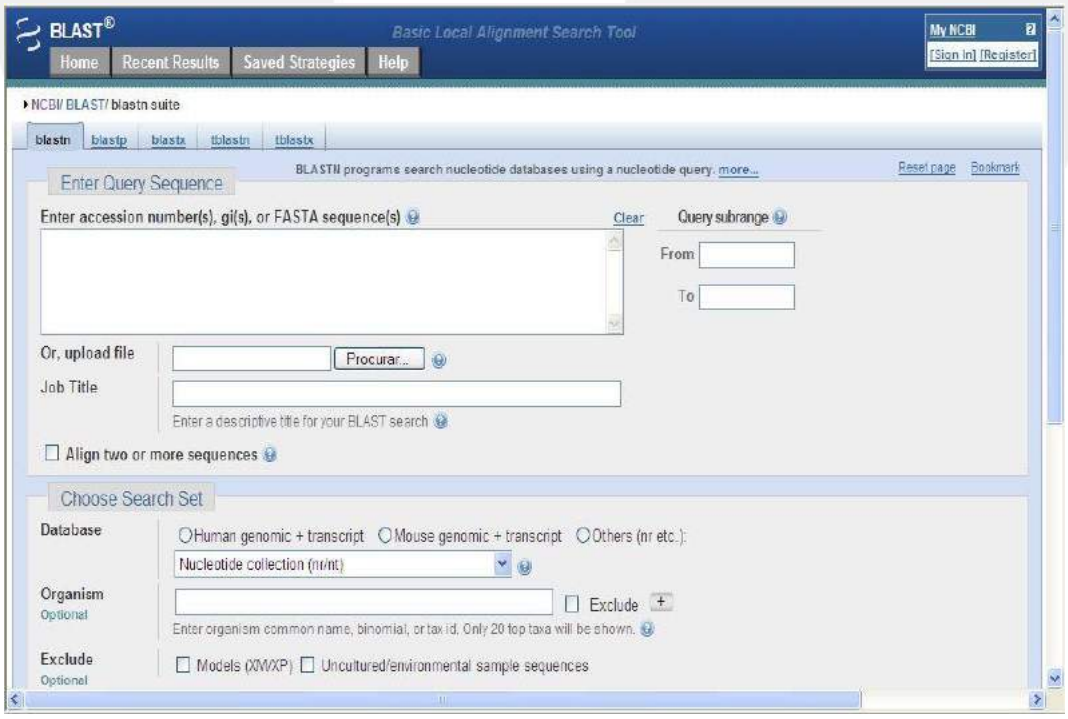
<http://www.ncbi.nlm.nih.gov/>.

FIGURA: PÁGINA INICIAL DO BLAST



<http://blast.ncbi.nlm.nih.gov/>.

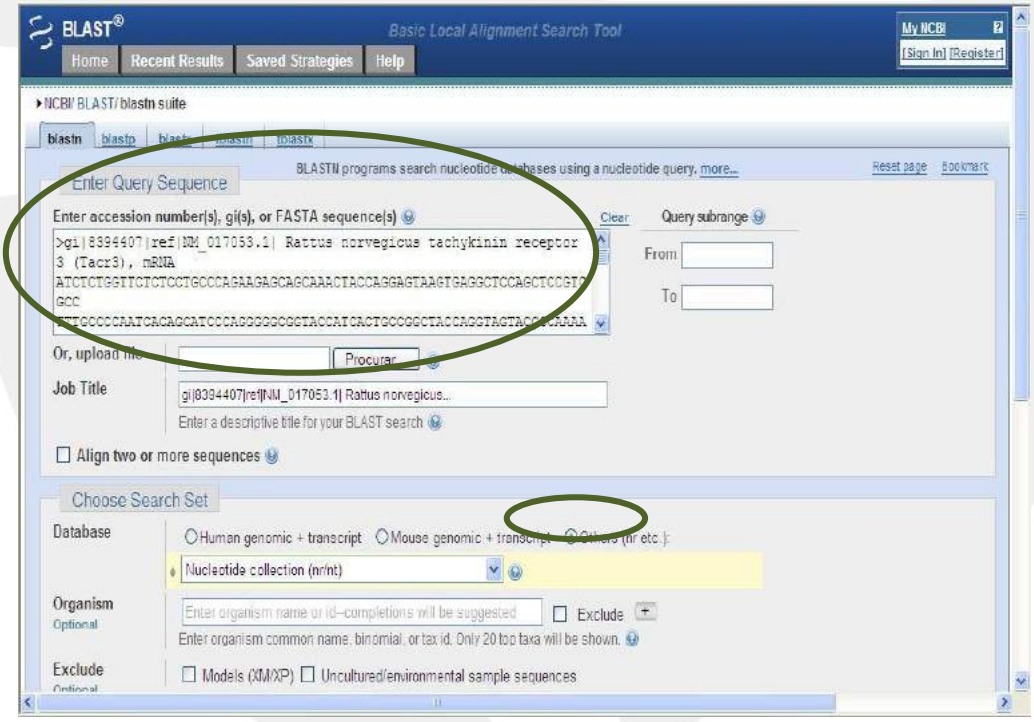
FIGURA: PÁGINA DO NUCLEOTIDE BLAST



http://blast.ncbi.nlm.nih.gov

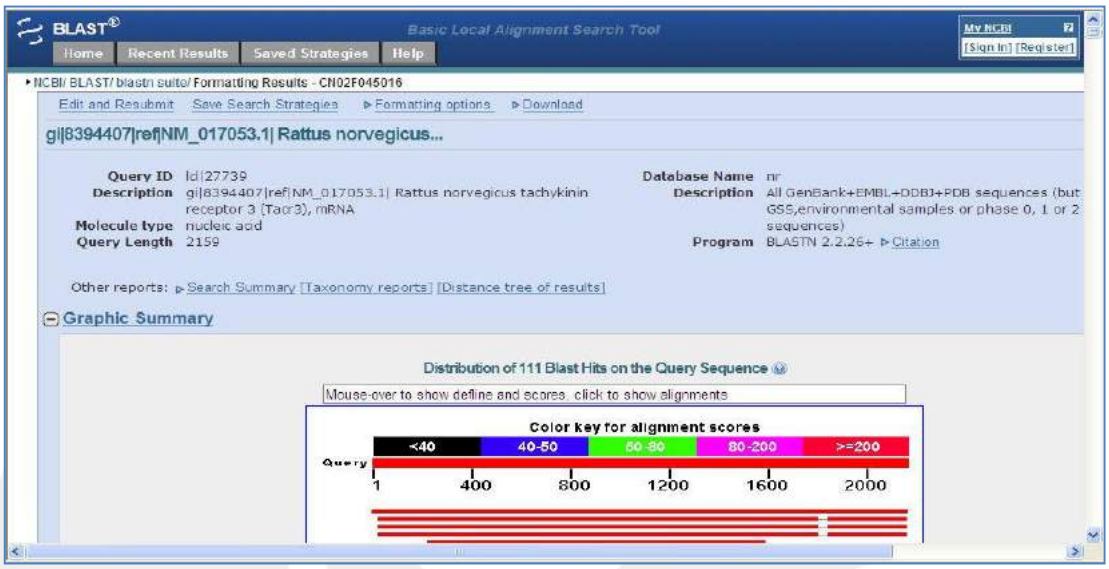
Depois de aberta a página do “nucleotide blast”, devemos colar a sequência de interesse em formato FASTA na caixa de texto disponível. Na opção databases selecionar a opção “others”, visto que estamos trabalhando com uma sequência de roedor (caso a sequência fosse de humano ou camundongo selecionaríamos as outras opções, respectivamente). Os demais parâmetros não devem ser alterados nesse primeiro momento. Abre-se, então, uma página bastante extensa contendo todos os resultados da análise do BLAST disponíveis.

FIGURA: PÁGINA DO “NUCLEOTIDE BLAST” CONTENDO A SEQUÊNCIA DO “RATTUS NORVEGICUS TACHYKININ RECEPTOR 3 (TACR3), MRNA” DE INTERESSE



<http://blast.ncbi.nlm.nih.gov/

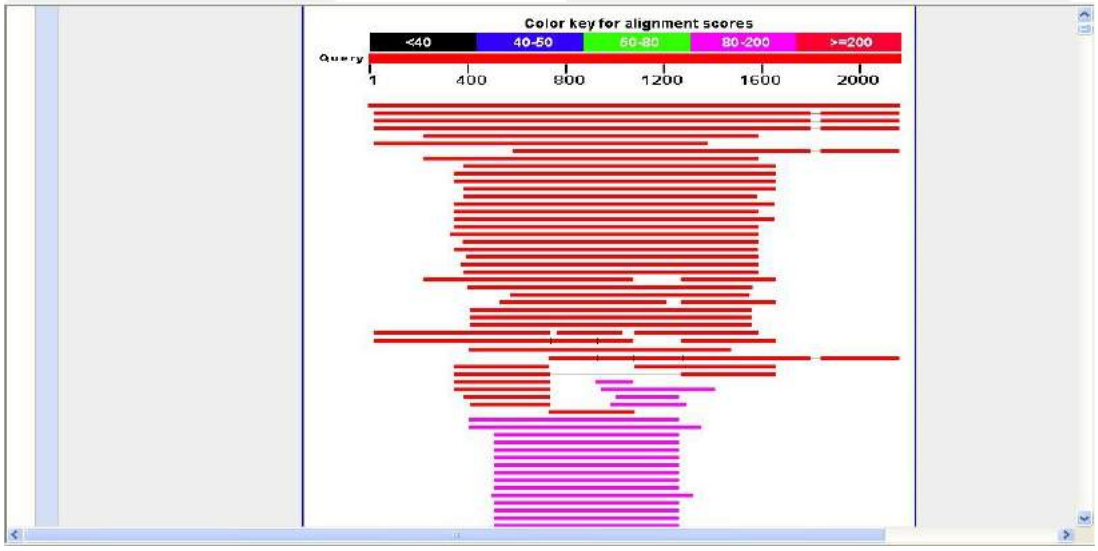
FIGURA: PÁGINA DE RESULTADOS DO “NUCLEOTIDE BLAST”



<http://blast.ncbi.nlm.nih.gov/Blast.cgi>.

Inicialmente devemos observar a representação esquemática dos resultados significativos. A barra vermelha representa a sequência de nucleotídeos da sequência que colamos anteriormente. As barras que estão abaixo das vermelhas representam esquematicamente as sequências de nucleotídeos que apresentam grau de homologia significativo (da mais para a menos significativa).

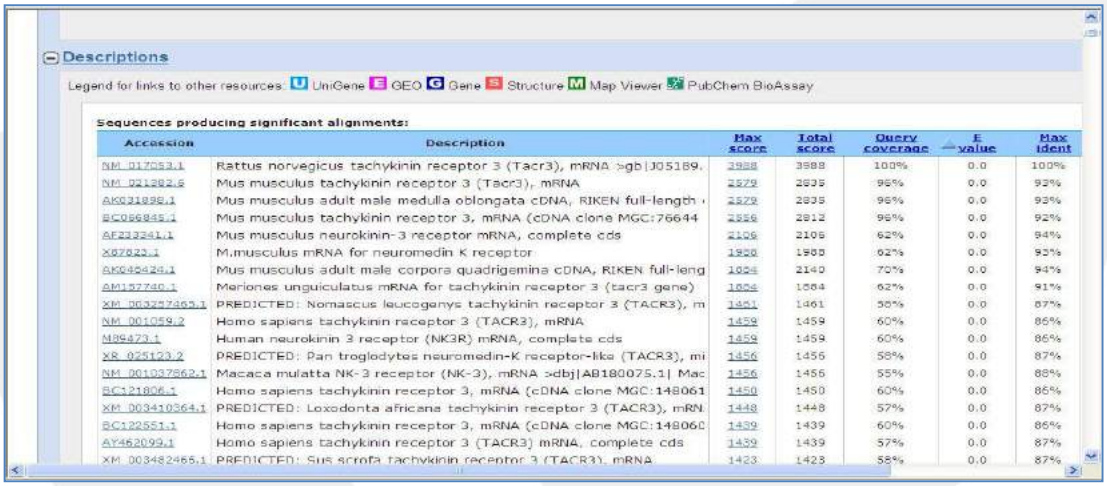
FIGURA: REPRESENTAÇÃO ESQUEMÁTICA DOS RESULTADOS SIGNIFICATIVOS



<http://blast.ncbi.nlm.nih.gov/Blast.cgi>.

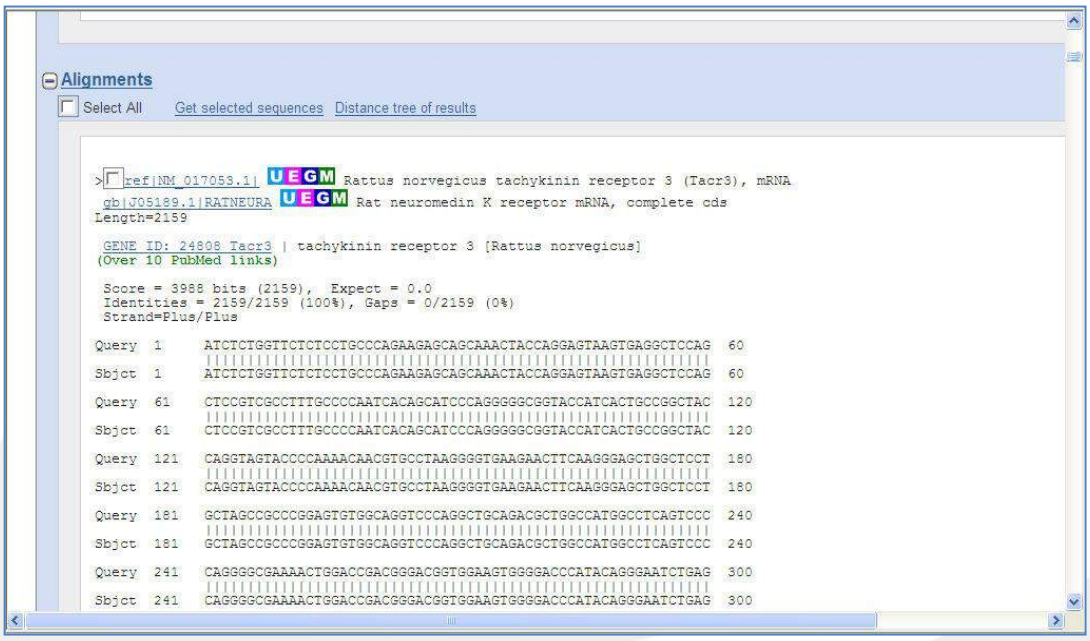
Abaixo das representações esquemáticas, segue-se uma lista com a descrição de todos os resultados obtidos com a análise do BLAST. Nessa lista, temos as descrições das sequências, e os valores de escore. Esses últimos fornecem informações sobre o grau de homologia entre a sequência em questão e a sequência de nosso interesse. Quanto maior o valor do escore, maior o grau de homologia entre as sequências. Outro dado interessante é valor “E”, o qual é um indicador do grau de significância da análise. Quanto menor esse valor, mais significativa é a análise. Abaixo dessa lista, segue os alinhamentos específicos de cada sequência analisada.

FIGURA: DESCRIÇÃO DOS RESULTADOS COM DESTAQUE PARA O SCORE E E VALUES



<http://blast.ncbi.nlm.nih.gov/Blast.cgi>.

FIGURA: DESCRIÇÃO DETALHADA DO ALINHAMENTO DE CADA UMA DAS SEQUÊNCIAS



<http://blast.ncbi.nlm.nih.gov/Blast.cgi>.

Na análise realizada, podemos perceber que a primeira sequência se refere a nossa própria sequência de interesse com um grau de homologia de 100%. Seguido a ela, vemos que a sequência do “Mus musculus tachykinin receptor 3 (TACR3) mRNA” apresenta o escore mais elevado, com uma similaridade de 96% e um E-value de 0, considerado excelente, e indicando que essa sequência tem um alto índice de similaridade com a nossa sequência de interesse. A partir desse dado, diversas análises com o objetivo de se estudar melhor as regiões de similaridade entre as sequências, com o intuito de avaliar possíveis relações evolutivas e funcionais podem ser realizadas.

O BLAST apresenta ainda diversas ferramentas para análise de primers, de proteínas, de domínios conservados, entre inúmeras outras análises disponíveis. Recomendo ler o tutorial do BLAST.

ANÁLISE DE PREDIÇÃO DOS SÍTIOS DE SPLICING

Diversos programas para avaliar sítios de splicing estão disponíveis. Daremos destaque aos seguintes programas:

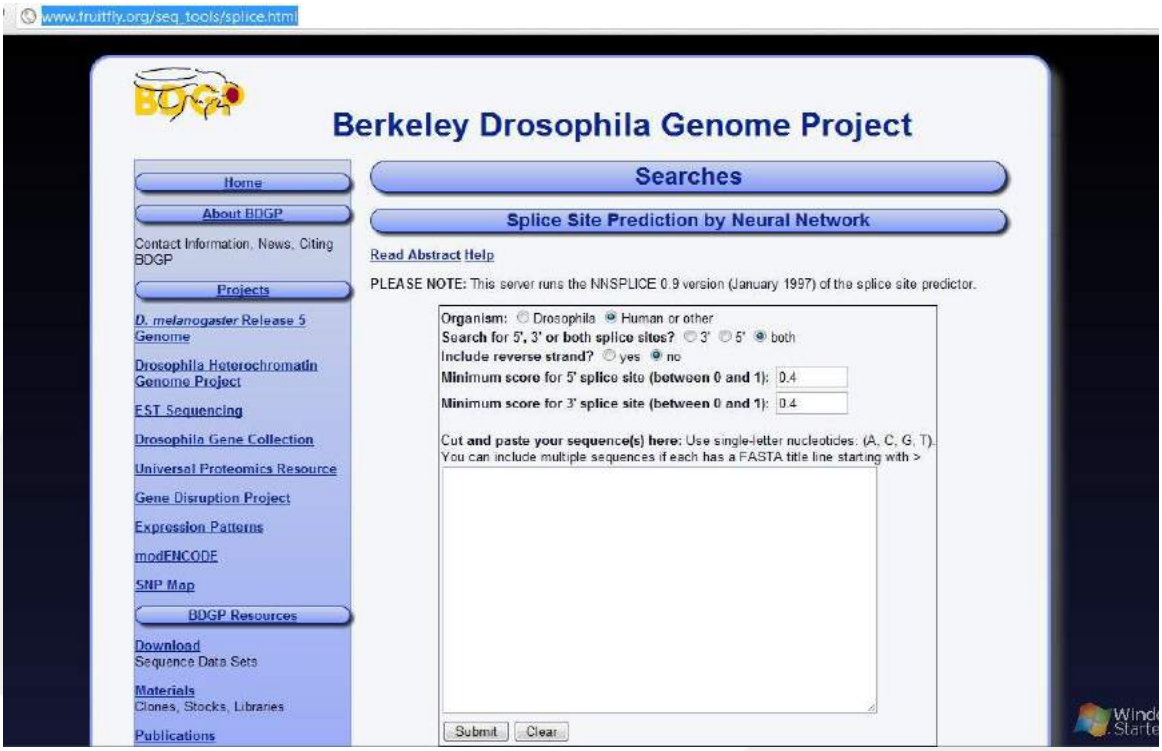
- NNSPLICE 0.9 disponível no endereço eletrônico http://www.fruitfly.org/seq_tools/splice.html para avaliação de sítios aceptores e doadores de splicing alternativo.
- RESCUE-ESE disponível no endereço eletrônico http://genes.mit.edu/burgelab/rescue-ese/ para avaliação de exonic splicing enhancer (ESEs) que podem apresentar importantes funções nos mecanismos de splicing alternativo e constitutivo. ESEs são pequenas sequências específicas de oligonucleotídeos que podem alterar os mecanismos de splicing alternativo presentes nos éxons.
- HUMAN SPLICING FINDER disponível no endereço eletrônico http://www.umd.be/HSF/ que agrega diversas matrizes com o intuito de entender melhor os diversos mecanismos de splicing existentes, apresentando inúmeras funções, tais como avaliação dos sítios crípticos de splicing, mas também análise de mutação, de múltiplos transcritos, entre outras funções.

NNSPLICE 0.9

Sítios de splicing são sequências sinais importantes para a determinação dos limites exônicos. Esse programa considera apenas genes que apresentam sítios consenso de splicing nas regiões 5’ (GT) e 3’ (AG). Agora, vamos ver na prática como funciona esse programa.

O primeiro passo é entrar no site http://www.fruitfly.org/seq_tools/splice.html. Nessa página inicial há a descrição do programa e a caixa de texto na qual devemos colocar nossa sequência de interesse, ou seja, àquela que queremos saber sobre os potenciais sítios doadores e aceptores de splicing. Destaco ainda os links “read abstract” e “help”, os quais ao serem clicados nos trazem detalhes sobre a teoria de como foi desenvolvido o programa.

FIGURA: PÁGINA INICIAL DO NNSPLICE 0.9

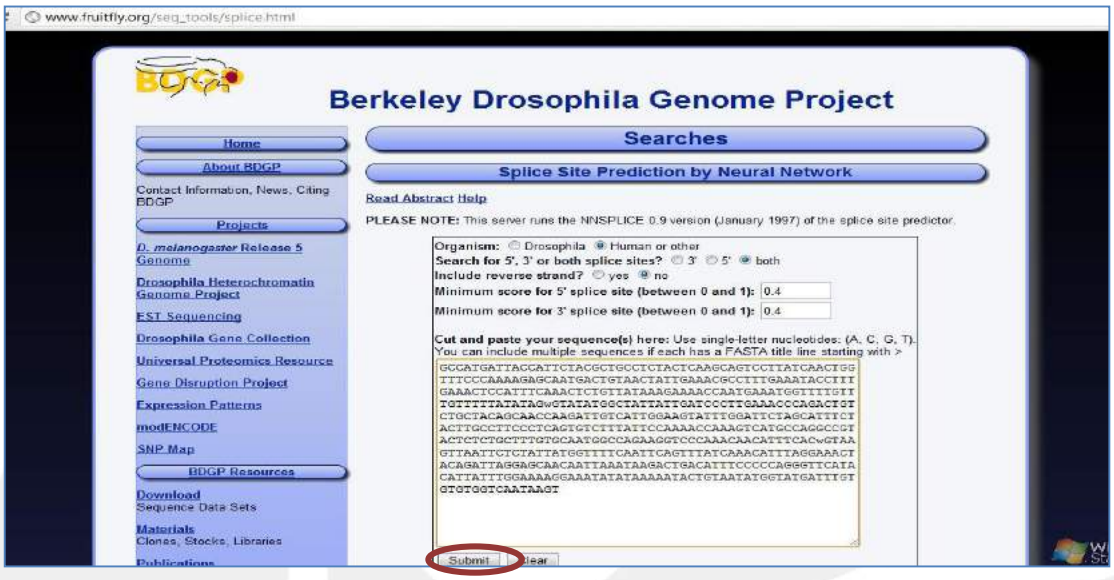


<http://www.fruitfly.org/seq_tools/splice.html>.

Como mencionado no item anterior, precisamos ter a nossa sequência de interesse em mãos para proceder à análise. Nesse caso, a minha sequência de interesse está localizada no gene TACR3 (copiei a parte da sequência que me interessava diretamente do site do NCBI, a qual foi explicada em detalhes. Com a sequência copiada, o passo seguinte é colá-la na caixa de texto.

Importante ressaltar que nesse caso específico não precisamos alterar os itens acima da caixa de texto, mas dependendo da pesquisa poderemos alterar ou não essas opções. Depois de colada a sequência deve-se clicar em “submit” para realizar a análise propriamente dita.

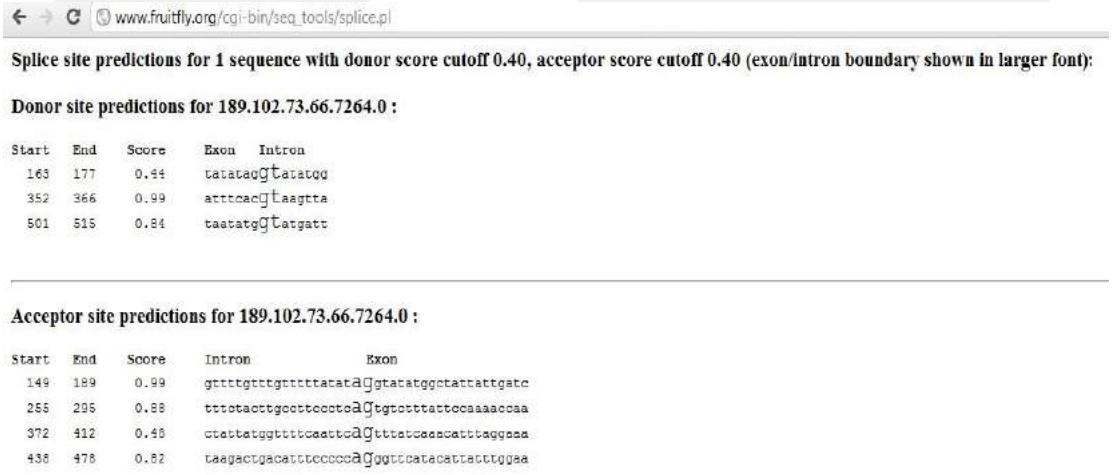
FIGURA: PÁGINA INICIAL DO NNSPLICE 0.9 COM A SEQUÊNCIA DE INTERESSE PRONTA PARA SER ANALISADA



<http://www.fruitfly.org/seq_tools/splice.html>.

A página de resultados é aberta nos trazendo uma série de informações que dependendo do objetivo da nossa pesquisa terão maior ou menor utilidade. Os sítios de splicing encontrados nessa sequência foram recuperados juntamente com seus escores. Os escores gerados pelo NNSPLICE variam de 0 a 1, sendo que um corresponde à combinação perfeita com o sítio consenso de splicing. Um limiar de 0.4 é usado pelo NNSPLICE para filtrar os resultados.

FIGURA: PÁGINA DE RESULTADOS DO NNSPLICE 0.9



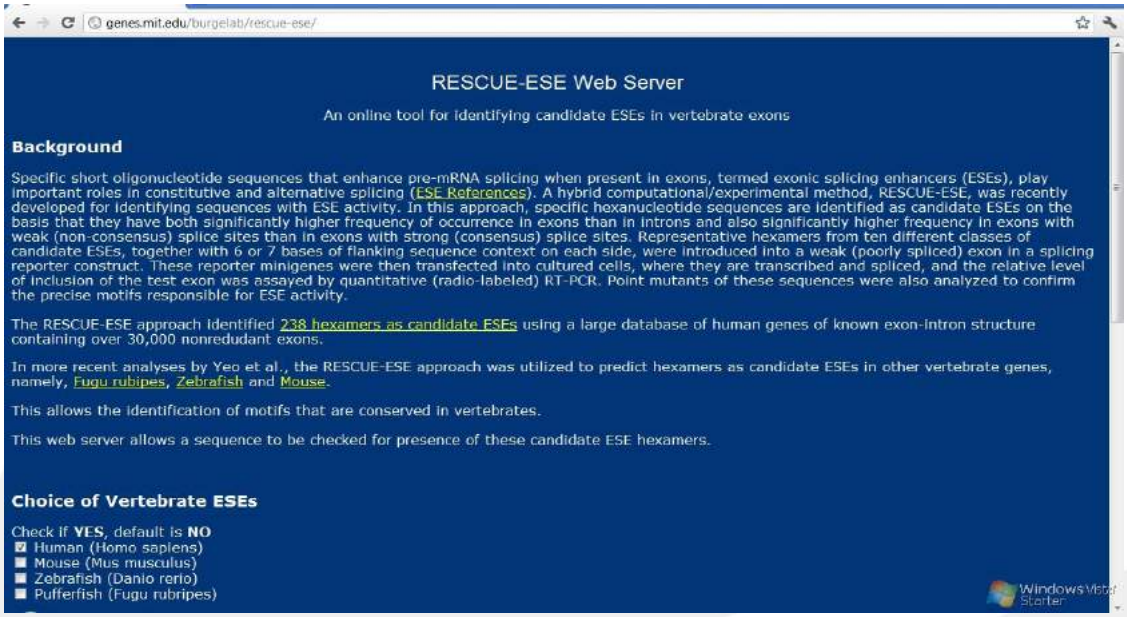
<http://www.fruitfly.org/cgi-bin/seq_tools/splice.pl>.

A partir da análise realizada, podemos inferir inúmeras outras análises comparando sequências selvagens e mutantes, a fim de avaliar se a presença de determinada mutação altera ou não determinado sítio de splicing. Nesse caso, devemos fazer um teste como o acima para a sequência selvagem e outro para a sequência mutante. Depois de realizadas as corridas, podemos comparar os valores de escores da região onde está localizada a mutação e ver se há alguma diferença entre o mutante e o selvagem. Lembrando que quando o limiar for menor que 0.4 para a sequência mutante podemos estar diante de uma quebra de sítio de splicing, e quando o escore for maior que 0.4 para o mutante, podemos estar diante da formação de um novo sítio de splicing. Obviamente, essas observações devem ser confirmadas em outros sites de bioinformática (recomenda-se no mínimo três) e sempre que possível realizada as análises experimentais das mesmas.

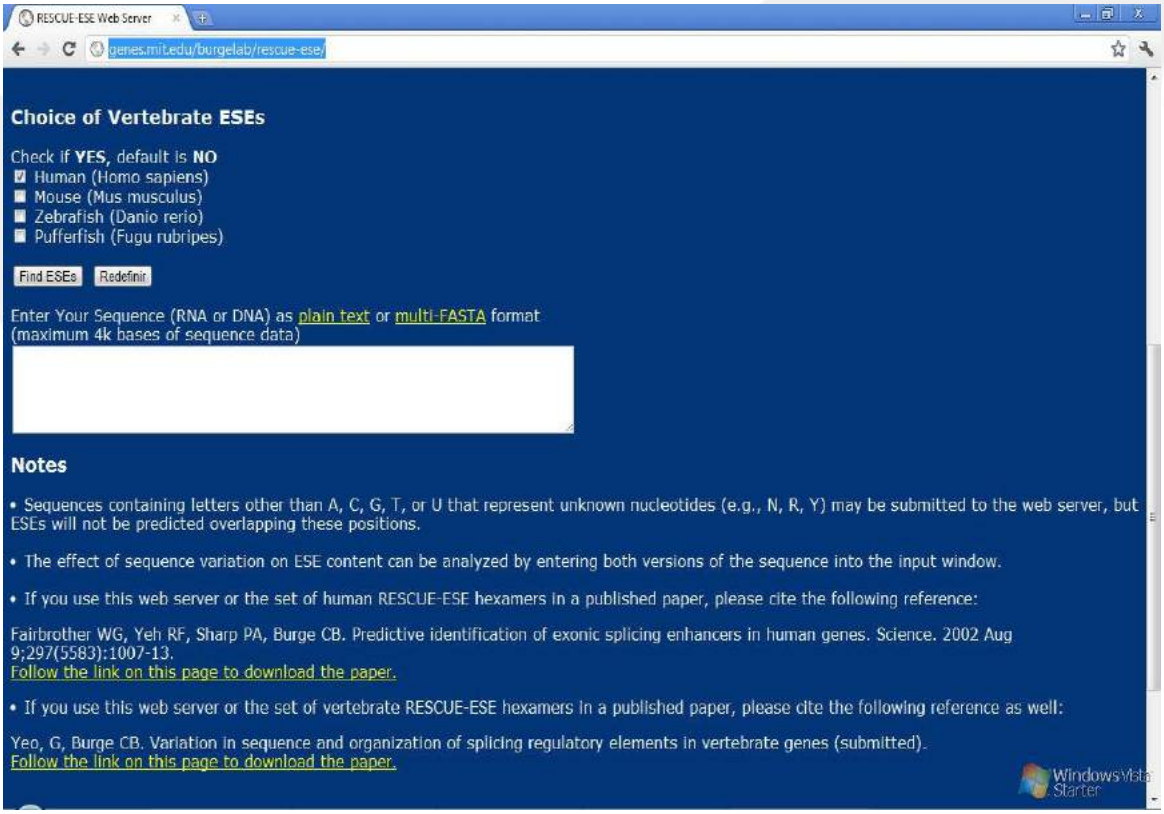
RESCUE-ESE

A página inicial do RESCUE-ESSE traz algumas informações sobre o conceito de ESEs, bem como a caixa de texto para a colagem da sequência de interesse e alguns links com referências importantes de serem lidas para quem vai trabalhar ativamente com o programa. Como citado anteriormente ESEs são pequenas sequências específicas de oligonucleotídeos que podem alterar os mecanismos de splicing alternativo presentes nos éxons. O RESCUE-ESE é um método híbrido (computacional/experimental) desenvolvido para identificar essas sequências ESEs, predizendo quais sequências têm atividade ESE.

FIGURA: PÁGINA INICIAL DO RESCUE-ESE.



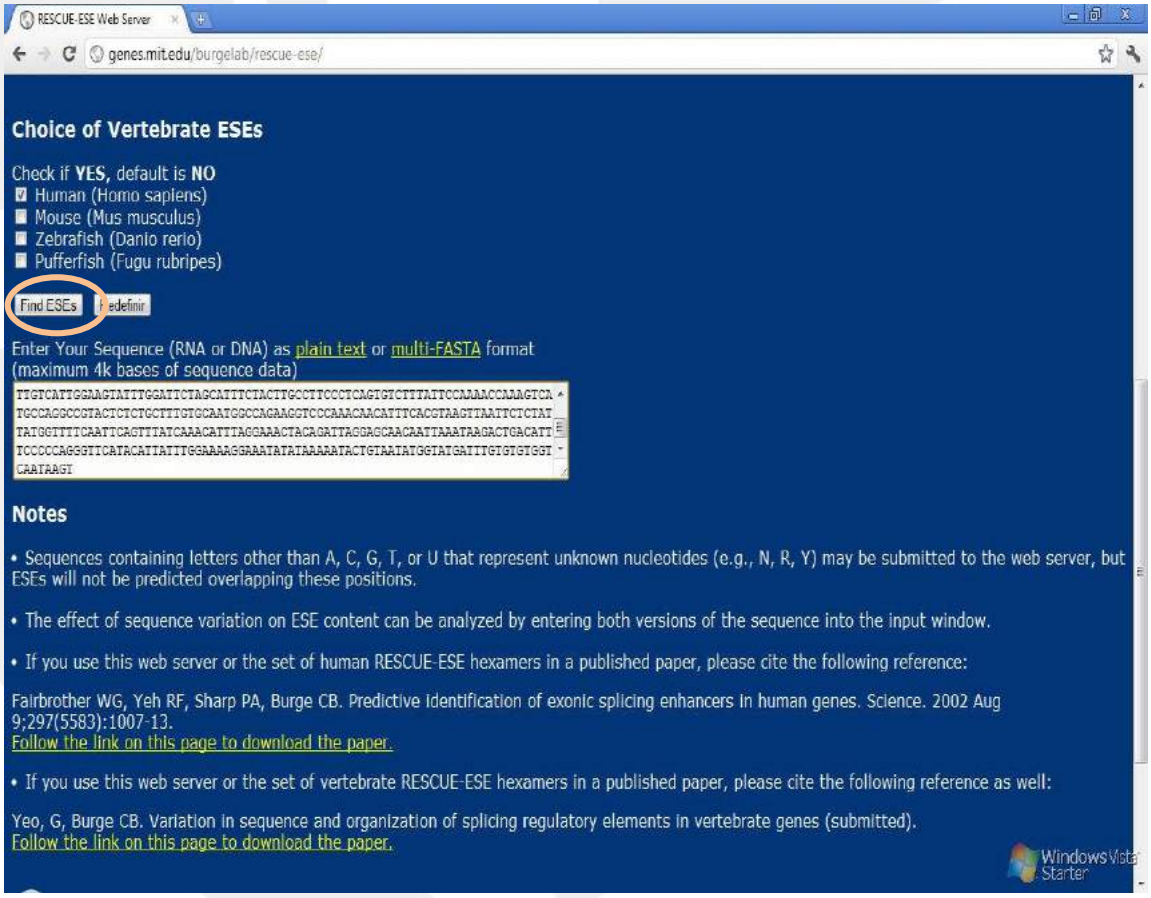
<http://genes.mit.edu/burgelab/rescue-ese/>.



<http://genes.mit.edu/burgelab/rescue-ese/>.

Para realizar esse tipo de análise, inicialmente devemos ter a sequência de interesse em mãos. Essa sequência deverá, então, ser colada na caixa de texto. Depois de colada e verificada a espécie que estamos analisando, nesse caso Homo sapiens, devemos clicar em “find ESEs” para realizar a análise propriamente dita, ou seja, encontrar as regiões na nossa sequência que tem atividade ESEs e, portanto podem alterar mecanismos de splicing.

FIGURA: PÁGINA INICIAL DO RESCUE-ESSE COM A SEQUÊNCIA DE INTERESSE PRONTA PARA SER ANALISADA

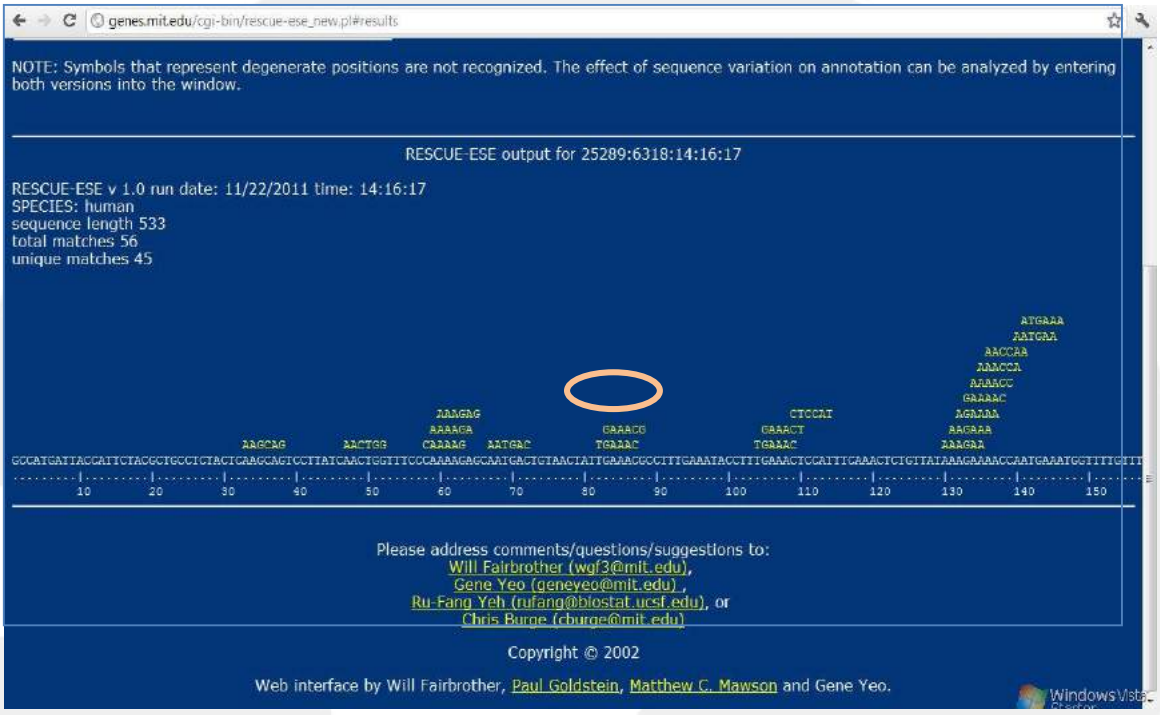


<http://genes.mit.edu/burgelab/rescue-ese/>.

A página de resultados gerada é na verdade uma representação esquemática de toda a sequência colada com os possíveis sítios candidatos ESEs marcados em amarelo. Da posse desses resultados, podemos analisar, por exemplo, se uma

determinada variante está localizada em um sítio de ESEs e a partir disso, realizar outras análises de bioinformática e experimentais voltadas para esse tipo de sequência gênica.

FIGURA: PÁGINA DE RESULTADOS DO RESCUE-ESSE



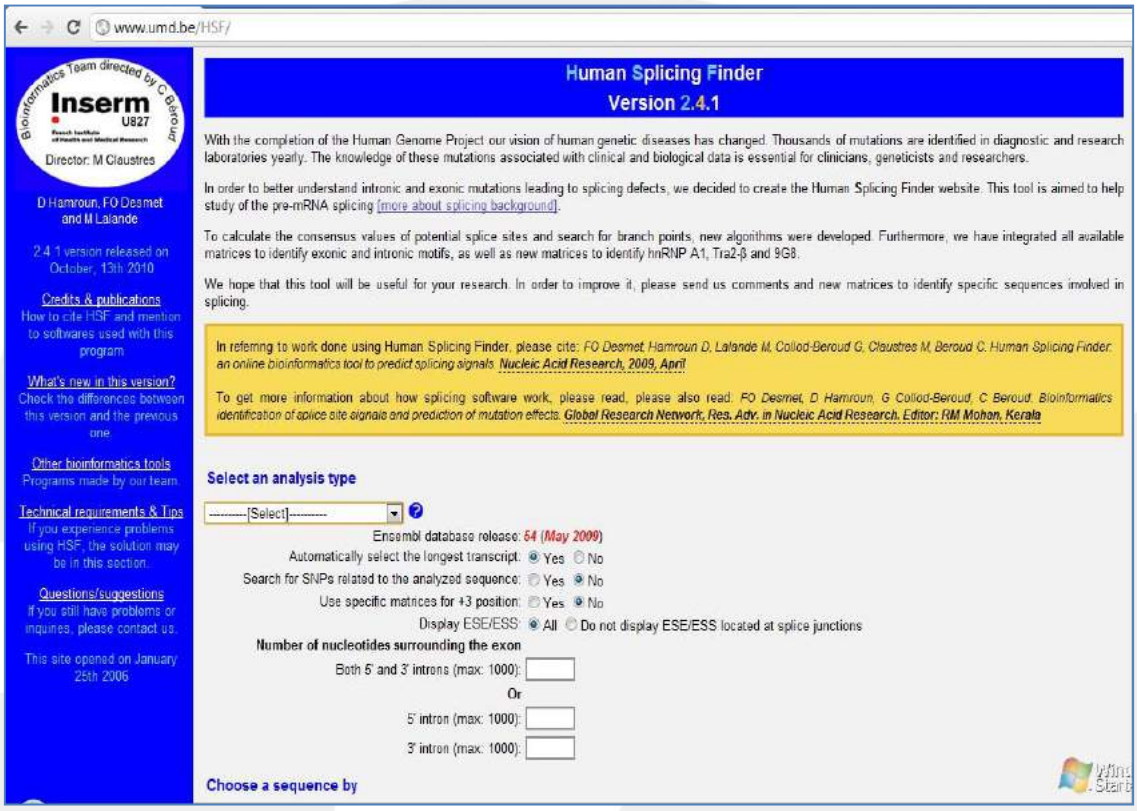
<http://genes.mit.edu/cgi-bin/rescue-ese_new.pl#results>.

HUMAN SPLICING FINDER

Todos os anos, milhares de mutações são identificadas. Entre essas, muitas afetam diretamente a expressão de proteínas, enquanto outras parecem influenciar os mecanismos de splicing do RNAm. O Human Splicing Finder (HSF) é uma ferramenta que visa facilitar a análise das diferentes mutações identificadas. Segue um exemplo de como usar o HSF.

Inicialmente devemos entrar no site do programa http://www.umd.be/HSF/. Essa página inicial apresenta o HSF, traz algumas valiosas referências para entendermos como o programa foi desenhado, e as ferramentas para a realização das análises propriamente ditas.

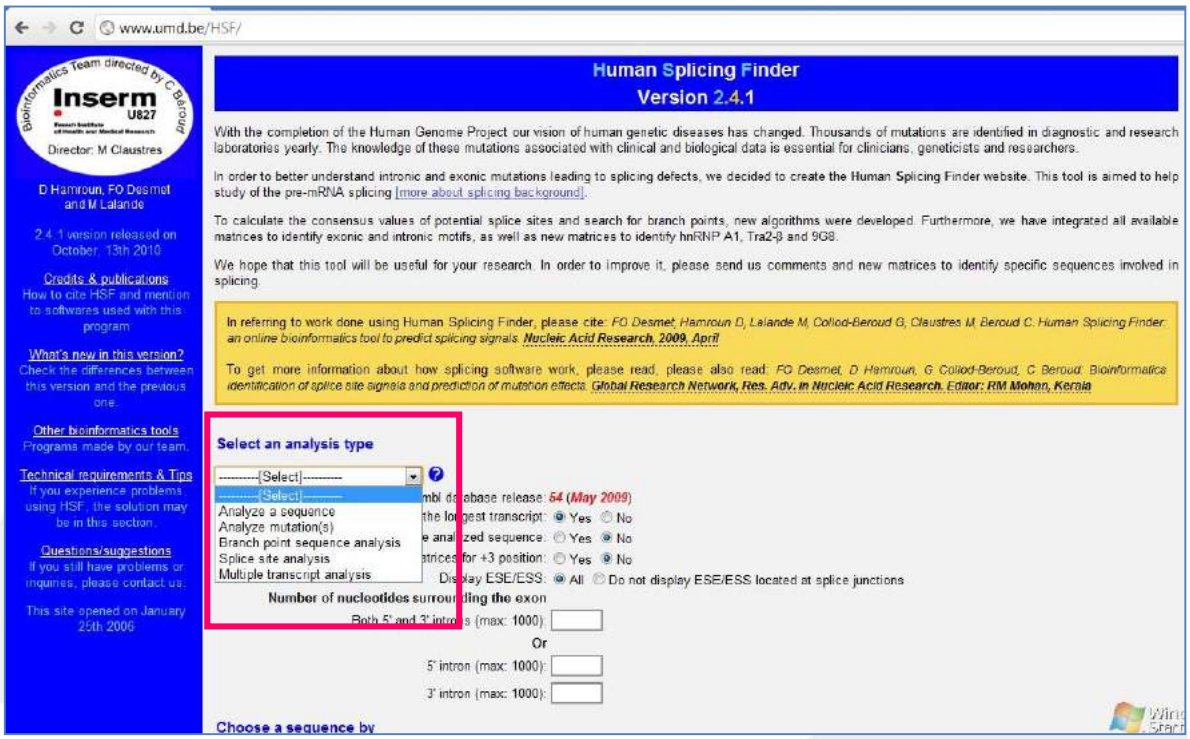
FIGURA: PÁGINA INICIAL DO HUMAN SPLICING FINDER



<http://www.umd.be/HSF/>

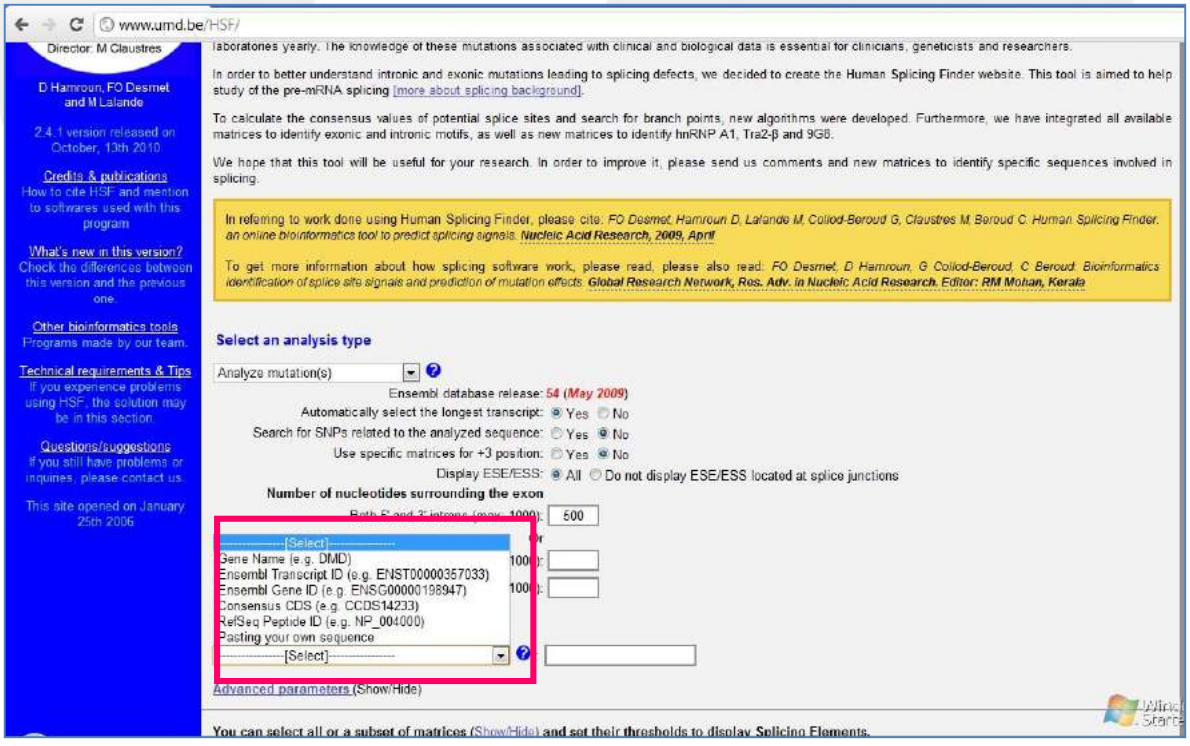
Nessa página inicial, é interessante observar que o HSF apresenta várias ferramentas de análises possíveis. Diante de cada caso específico, podemos selecionar a análise que melhor atende ao nosso propósito. Além disso, podemos selecionar entre as várias opções como a sequência será selecionada, se pelo nome do gene, do transcrito, peptídeo, entre outras formas de seleção.

FIGURA: PÁGINA INICIAL DO HUMAN SPLICING FINDER COM DESTAQUE PARA AS ANÁLISES PASSÍVEIS DE SEREM REALIZADAS NESSE PROGRAMA



<http://www.umd.be/HSF/>

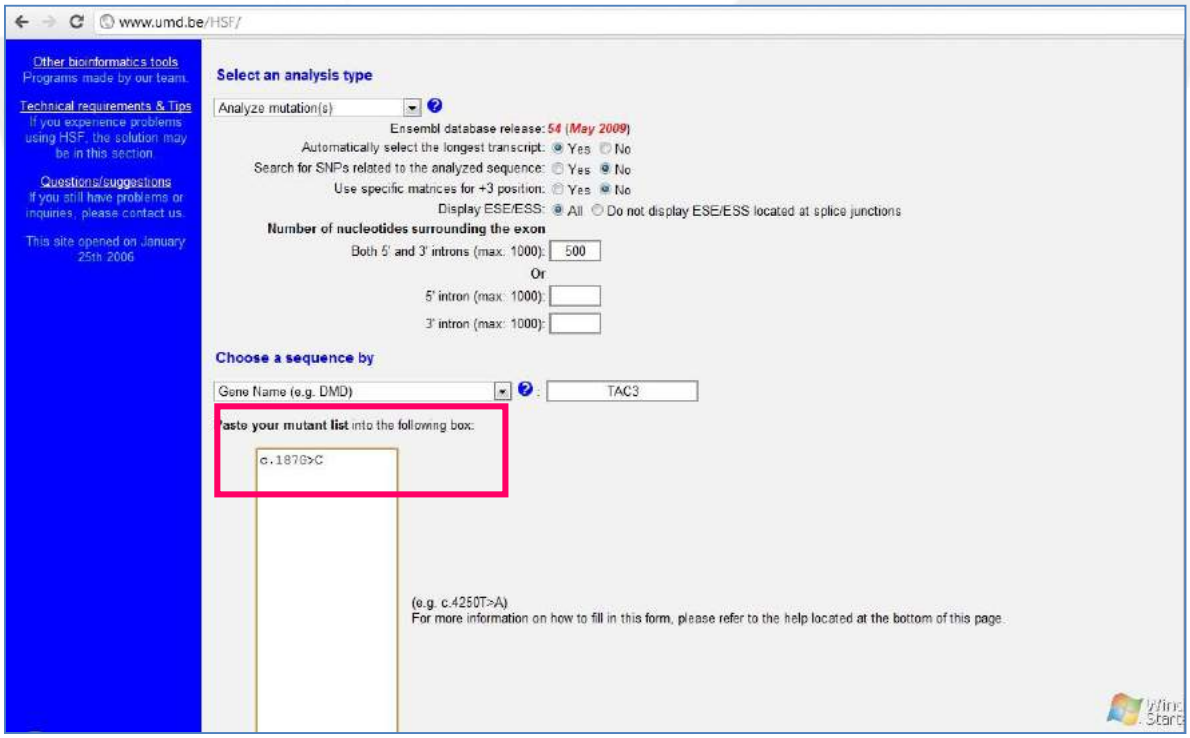
FIGURA: PÁGINA INICIAL DO HUMAN SPLICING FINDER COM DESTAQUE PARA AS DIFERENTES FORMAS DE SELEÇÃO DA SEQUÊNCIA DE INTERESSE



<http://www.umd.be/HSF/>.

No nosso exemplo, vamos fazer uma análise bastante comum nos laboratórios de pesquisa científica, que é a análise de mutações. Com esse tipo de análise, queremos responder a pergunta se a alteração que encontramos altera ou não algum mecanismo de splicing. Dessa forma, considerando a alteração do gene TAC3, c.187G>C ou p. A63P, vamos selecionar a opção, “analyse mutations”. Após essa seleção colocaremos como número de nucleotídeos ao redor das extremidades 5’ e 3’ do éxon cerca de 500 nucleotídeos (o número varia dependendo da análise), e selecionaremos a sequência pelo nome do gene (outras opções também estão disponíveis e dependem da preferência de cada pesquisador), e abaixo colocaremos a nossa variante em questão (c.187G>C), e procederemos a análise.

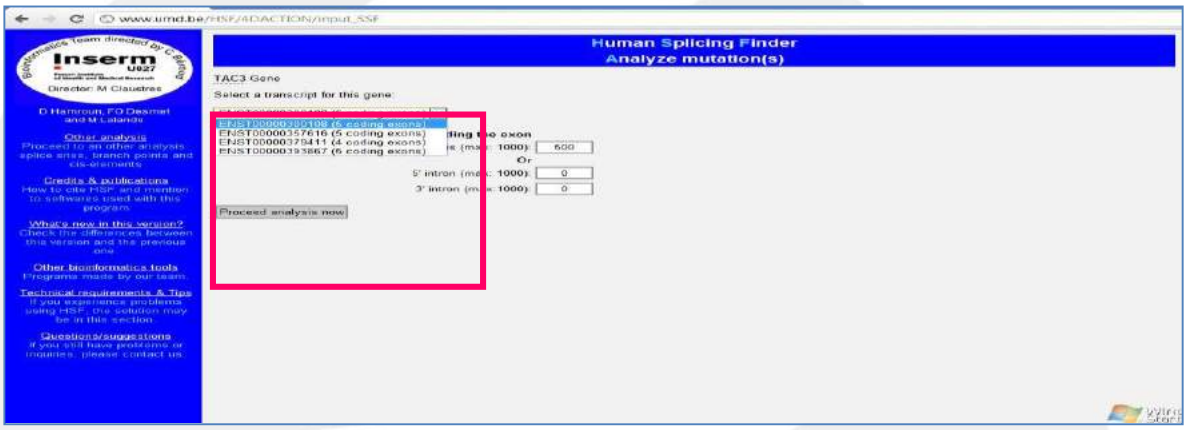
FIGURA: PÁGINA INICIAL DO HUMAN SPLICING FINDER COM TODOS OS DADOS DE INTERESSE PARA A ANÁLISE PREENCHIDOS



<http://www.umd.be/HSF/>

Importante ressaltar que caso o gene em questão tenha mais de um transcrito, outra janela se abrirá, solicitando que selecionemos o transcrito de interesse. Essa informação é variável e depende de cada trabalho em particular.

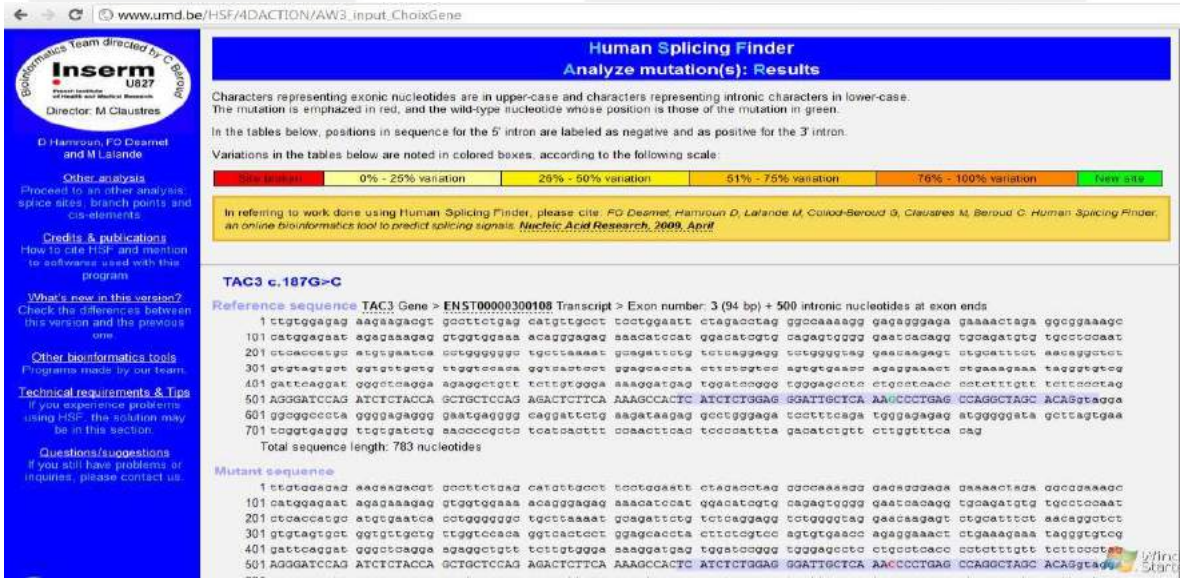
FIGURA: PÁGINA DO HSF SOLICITANDO A SELEÇÃO DO TRANSCRITO DE INTERESSE



<http://www.umd.be

Depois de realizada a análise, será gerada uma página de resultados. Nessa página, teremos inicialmente a nossa sequência selvagem e mutante, e abaixo dessas sequências, as análises realizadas.

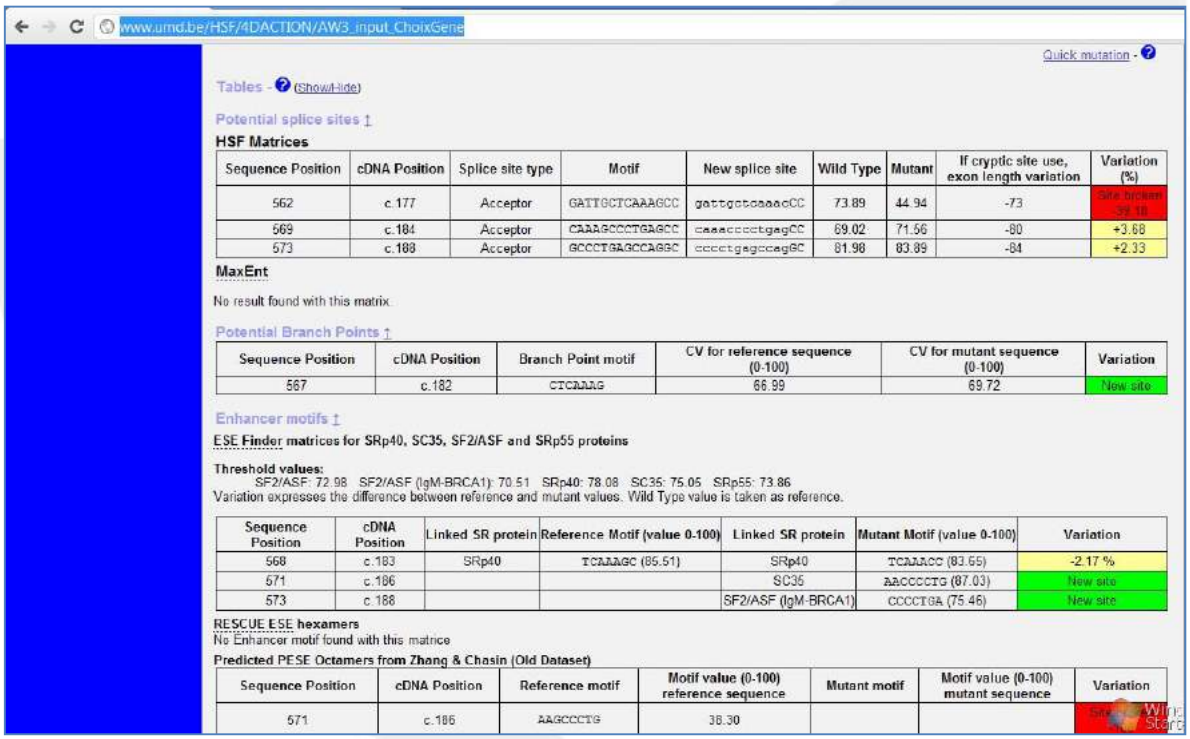
FIGURA: PÁGINA DE RESULTADOS DO HSF



<http://www.umd.be/HSF/4DACTION/AW3_input_ChoixGene>.

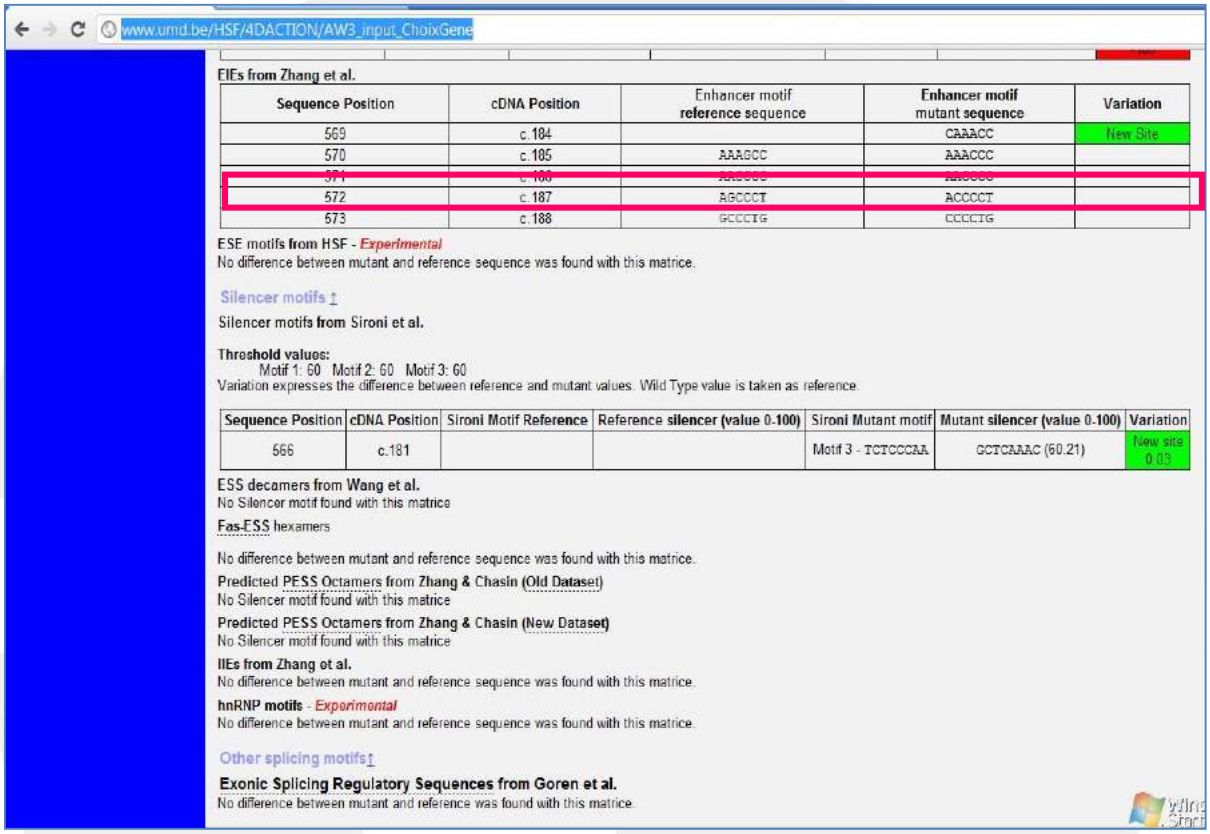
Para realizar a análise dos resultados propriamente dita, devemos primeiramente encontrar a variante em questão pela posição da troca nucleotídica nas tabelas apresentadas abaixo das sequências. No nosso exemplo, a alteração está na posição 187, devemos então procurar se há indicação dessa posição nos resultados apresentados, e analisar se ocorre alguma possível alteração nos sítios de splicing, tais como criação de um novo sítio, quebra de um sítio, entre outras. No nosso exemplo, a alteração em questão não altera os mecanismos de splicing alternativo analisados por esse programa.

FIGURA: PÁGINA DE RESULTADOS DO HSF



http://www.umd.be/HSF/4DACTION/AW3_input_ChoixGene

FIGURA: PÁGINA DE RESULTADOS DO HSF COM DESTAQUE PARA A POSIÇÃO DA TROCA NUCLEOTÍDICA DE INTERESSE



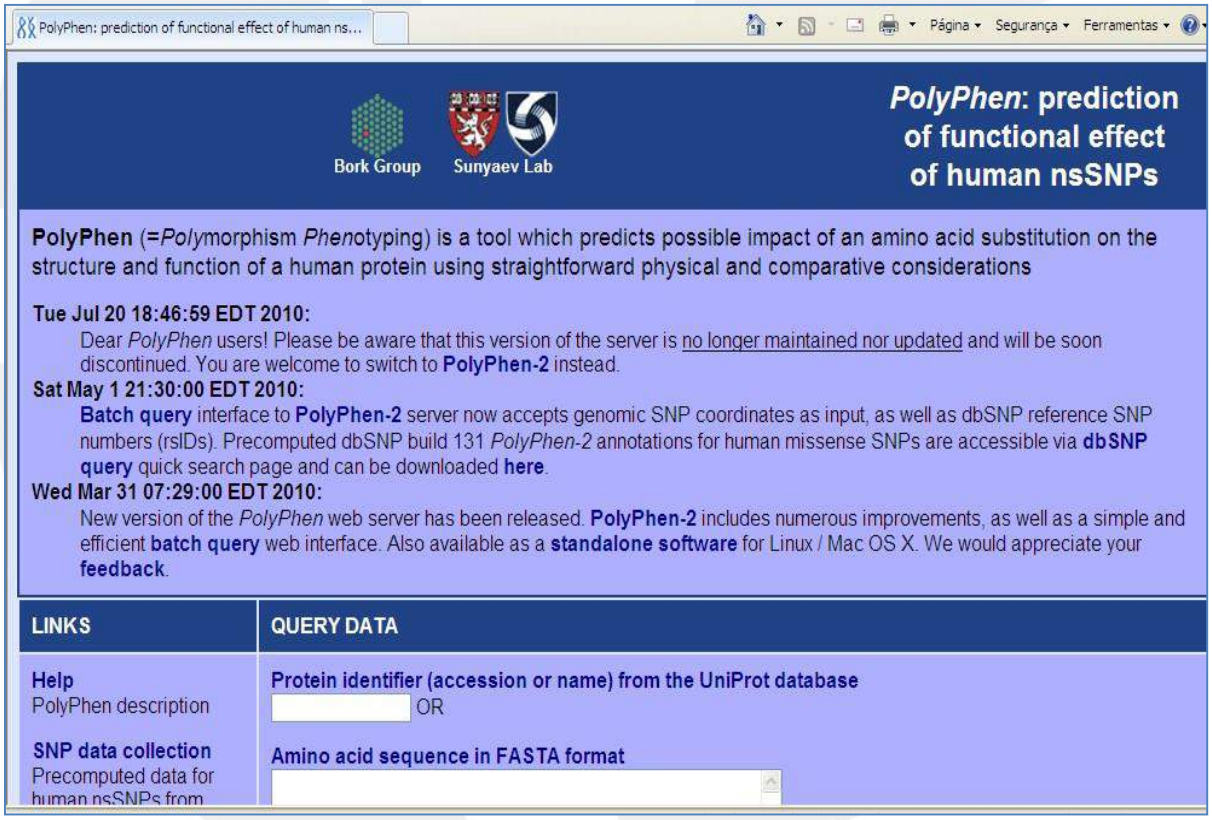
<http://www.umd.be/HSF/4DACTION/AW3_input_ChoixGene>.

Inúmeros outros programas para avaliar diversos mecanismos de splicing estão disponíveis. Nas referências bibliográficas, estão listados os mais diversos programas para predição de sítios de splicing, além de artigos explicativos sobre o assunto. A apresentação de programas para analisar os efeitos que novas variantes possam ter sobre a estrutura proteica final. Daremos ênfase ao programa PolyPhen e SIFT.

POLYPHEN

O PolyPhen (Polymorphism Phenotyping) é uma ferramenta de bioinformática que prediz os possíveis impactos que a substituição de aminoácidos podem ter sobre a estrutura e função das proteínas, utilizando para isso diversos algoritmos. A página inicial do PolyPhen apresenta uma introdução geral sobre o programa, bem como as ferramentas de análise propriamente ditas.

FIGURA: PÁGINA INICIAL DO POLYPHEN

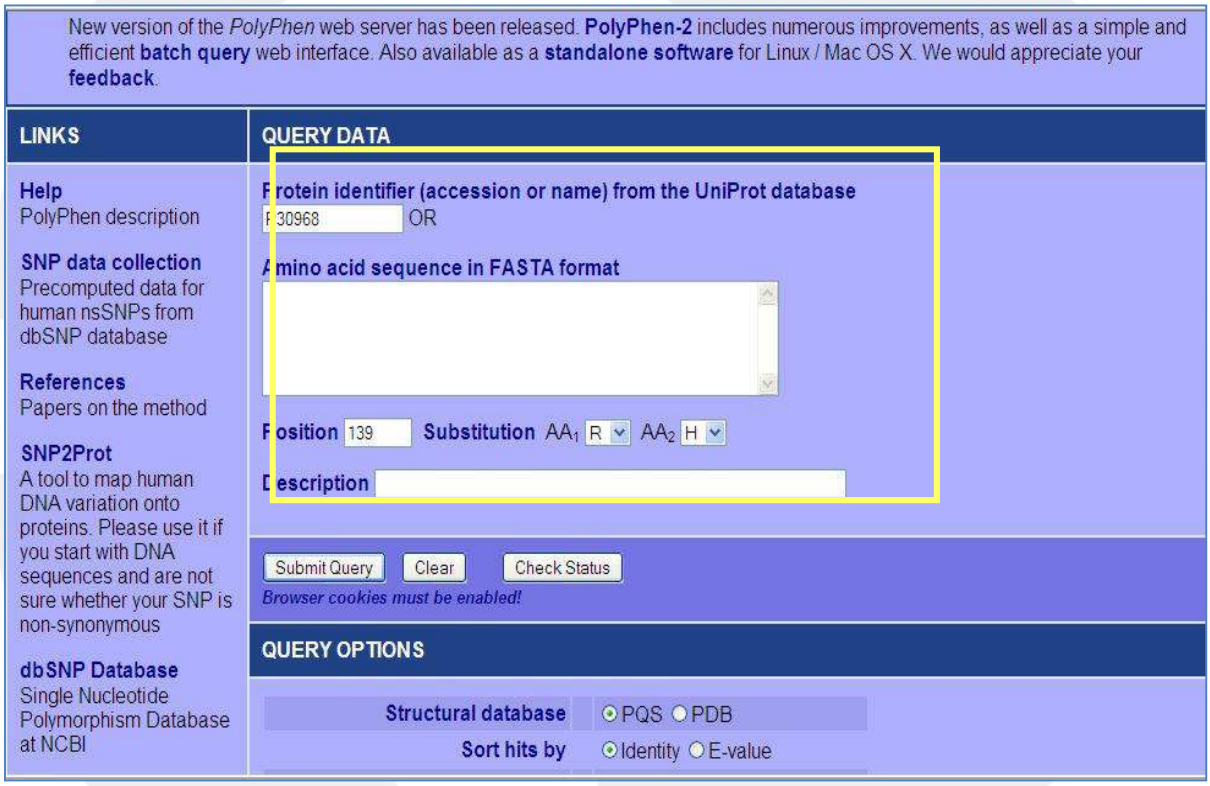


http://genetics.bwh.harvard.edu/pph

Para exemplificar o uso do PolyPhen, vamos considerar a mutação p.R139H da proteína GNRHR. Apenas com esses dois dados, eu posso realizar algumas análises in silico sobre o efeito dessa mutação sobre a estrutura proteica. Inicialmente, devemos escrever o nome da proteína de acordo com a base de dados do UniProt no campo “Protein Identifier”. Abaixo devemos colocar a posição da

nossa variante, ou seja, 139 na proteína madura, bem como a troca de aminoácido que está presente. Depois de preenchido esses dados, podemos submeter à análise. Nesse caso, não precisamos alterar os demais parâmetros de análise.

FIGURA: PÁGINA INICIAL DO POLYPHEN COM OS DADOS DE INTERESSE PREENCHIDOS



<http://genetics.bwh.harvard.edu/pph>

Enquanto a análise está sendo executada, abre-se uma nova página indicando essa ocorrência.

FIGURA: PÁGINA DO POLYPHEN INDICANDO QUE A ANÁLISE ESTÁ SENDO PROCESSADA

<http://genetics.bwh.harvard.edu/cgi-bin/ggi/ggi.cgi>.

Após a análise ser efetuada, devemos clicar em “view” e será aberta a página de resultados gerada pela análise no PolyPhen.

FIGURA: PÁGINA DE RESULTADOS DO POLYPHEN

<http://genetics.bwh.harvard.edu

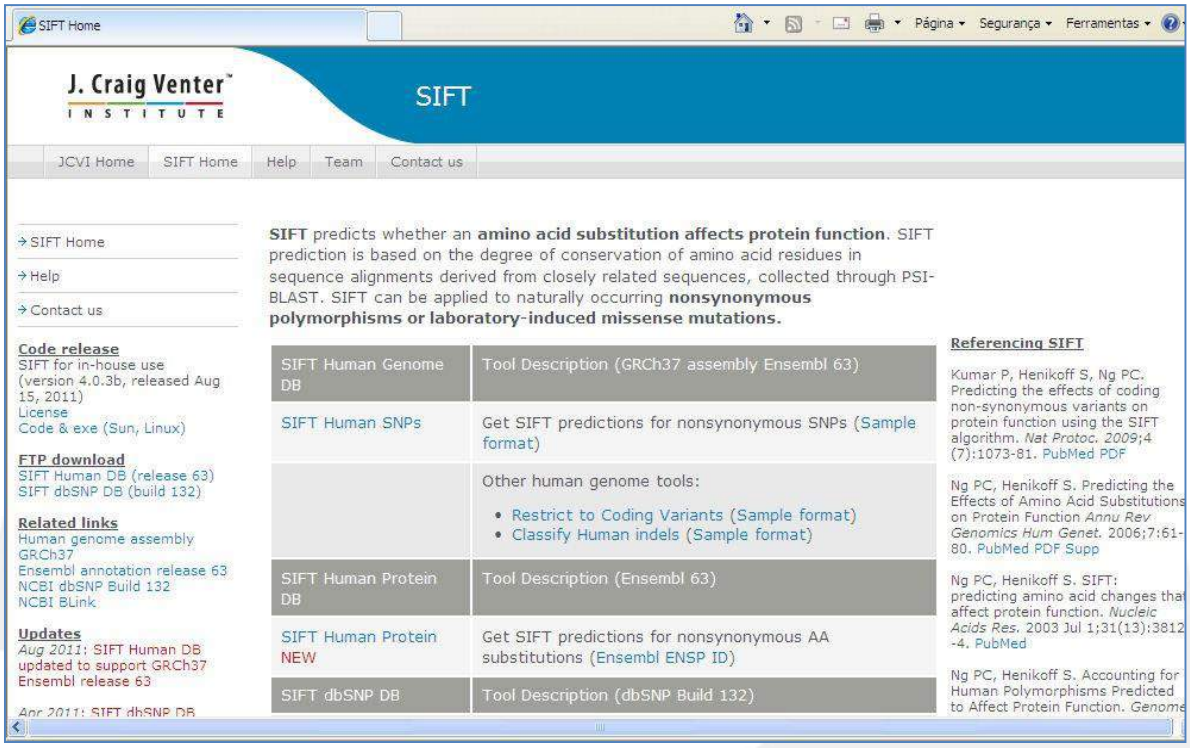
No alto da página de resultados, temos a descrição da variante propriamente dita, ou seja, o nome da variante de acordo com o UniProt, a posição da troca de aminoácidos, e os aminoácidos que são substituídos. Abaixo dessa descrição, temos a análise de predição propriamente dita. Nesse caso específico, o programa nos mostra que a troca de uma arginina por uma histidina na posição 139 do GNRHR provavelmente tem um efeito danoso para a proteína.

Obviamente, essa análise não é conclusiva e deve ser confirmada em outros sites de predição, bem como por meio de uma análise experimental mais aguçada. Outros possíveis resultados que poderiam ser encontrados é a possibilidade da troca de aminoácidos ser benigna, ou seja, não alterar significativamente a função da proteína; ou ainda, poderíamos ter como resultado que a variante em questão possivelmente é danosa para a proteína, ou seja, há possibilidade de ter efeito funcional, o qual precisa ser melhor estudado. A diferença dessa última alternativa para a encontrada no nosso exemplo, é que a primeira tem mais chances de realmente ter um efeito funcional sobre a proteína, enquanto a segunda é apenas uma possibilidade devido à troca de aminoácidos.

SIFT

Outro programa interessante para análise de substituição de aminoácidos é o SIFT. Esse programa é baseado no grau de conservação dos resíduos de aminoácidos no alinhamento de sequências. A diferença para o PolyPhen é que o SIFT apresenta variadas ferramentas de análises, as quais podem ser escolhidas de acordo com o objetivo do pesquisador na página inicial do site.

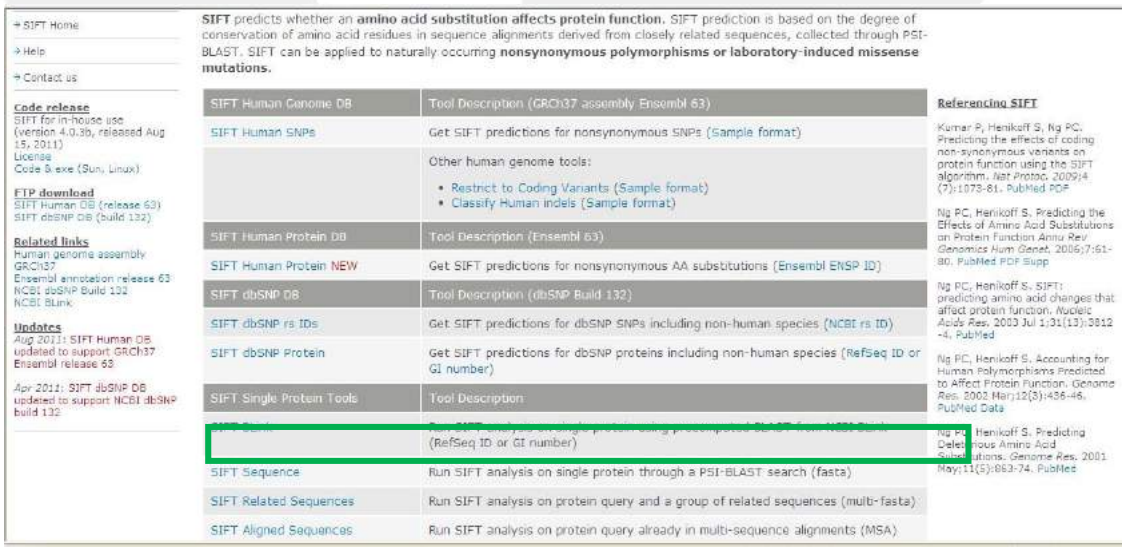
FIGURA: PÁGINA INICIAL DO SIFT



<http://sift.jcvi.org/>.

Analisando essa página, vamos encontrar uma série de possibilidades de análise. Vamos utilizar como exemplo o link “SIFT SEQUENCE”.

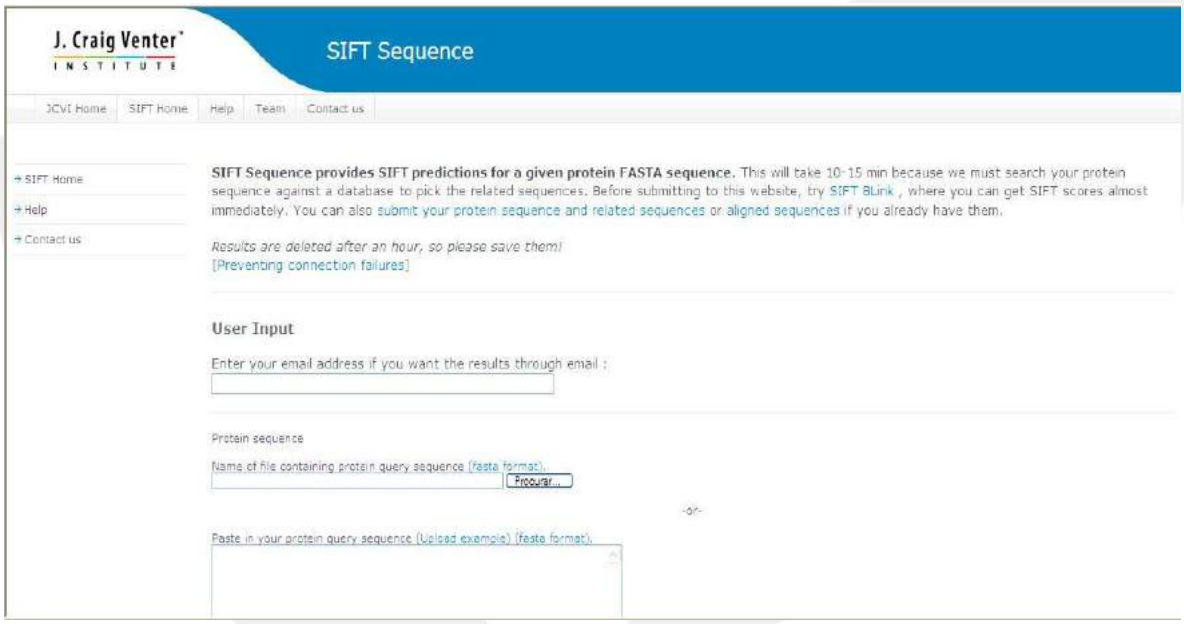
FIGURA: PÁGINA INICIAL DO SIFT COM DESTAQUE PARA O LINK “SIFT SEQUENCE”



FONTE: Disponível em: <http://sift.jcvi.org/>. Acesso em: 07 Abr. 2014.

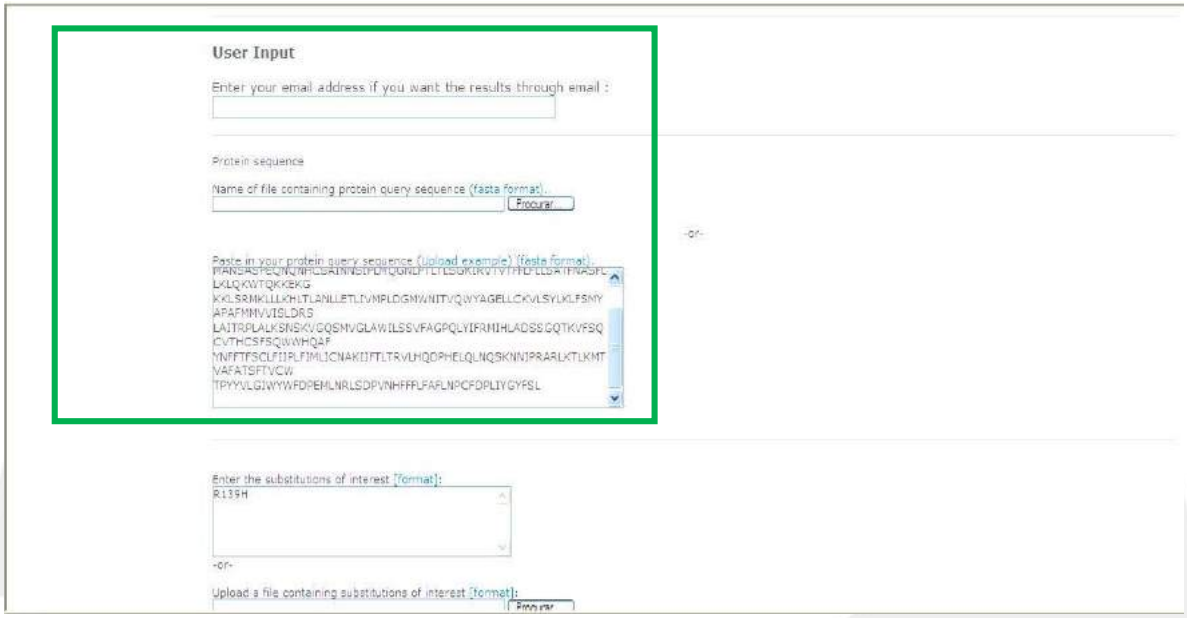
Quando clicarmos no link de interesse, será aberta uma nova página, na qual colocaremos os dados que queremos analisar. É interessante observar que nessa página, podemos colocar um endereço de e-mail para que a análise seja enviada diretamente para esse e-mail. Além disso, alguns detalhes devem ser levados em consideração antes de preenchermos os dados solicitados. O primeiro é que a sequência de aminoácidos a ser colada deve estar obrigatoriamente em formato FASTA. Para pegarmos a sequência nesse formato devemos entrar no NCBI e na base de dados de proteína, procurar pela proteína de interesse em formato fasta. Abaixo dessa caixa de texto, devemos colocar a variante de interesse, a qual se deseja saber se tem algum efeito sobre a estrutura proteica final ou não.

FIGURA: PÁGINA DO SIFT SEQUENCE



<http://sift.jcvi.org/www/SIFT_seq_submit2.html>.

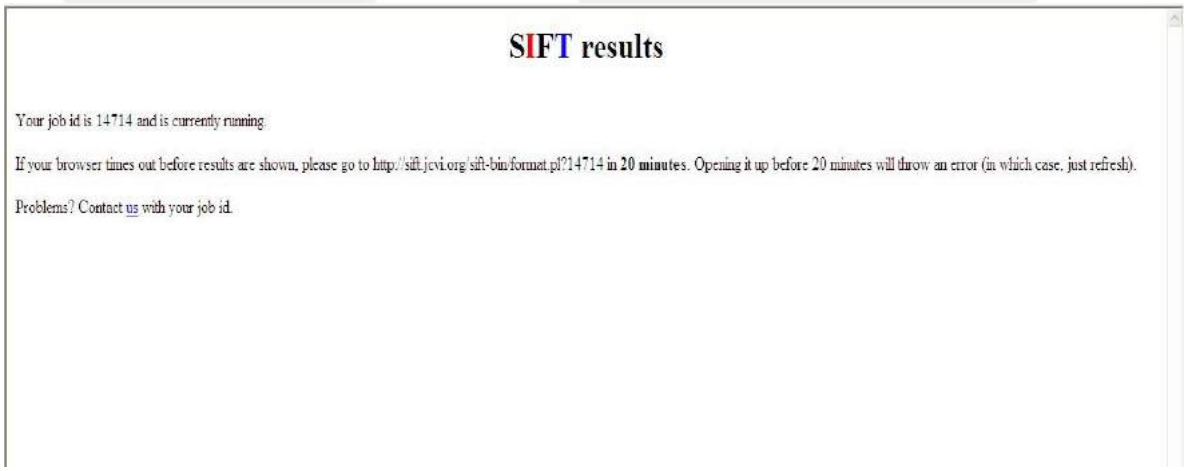
FIGURA: PÁGINA DO SIFT SEQUENCE COM OS DADOS DE INTERESSE PREENCHIDOS



<http://sift.jcvi.org/www/SIFT_seq_submit2.html>.

Após os dados serem preenchidos, podemos enviar a consulta. Uma nova página será aberta indicando que a análise está sendo realizada. Importante ressaltar que essa análise pode demorar algum tempo para ser completa.

FIGURA: PÁGINA DO SIFT SEQUENCE INDICANDO QUE A ANÁLISE ESTÁ SENDO PROCESSADA



<http://sift.jcvi.org/sift-bin/SIFT_seq_submit2.pl>.

Depois de realizada a análise, uma série de resultados será gerada pelo programa.

FIGURA: PÁGINA COM OS DIVERSOS RESULTADOS GERADOS PELA ANÁLISE DO SIFT SEQUENCE

SIFT results

Your job id is 14714 and is currently running.

If your browser times out before results are shown, please go to http://sift.jcvi.org/sift-bin/format.pl?14714 in **20 minutes**. Opening it up before 20 minutes will throw an error (in which case, just refresh).

Problems? Contact us with your job id.

Your results will be available here for the next 24 hours.
53 sequences were selected to be closely related to your query sequence.
PSIBLAST alignment of submitted sequences
Alignment in FASTA format (for modification)
The alignment taken from PSIBLAST is returned in msf format.
Note: Xes are placeholders at the beginning and end of sequences. While - means a gap in the alignment an X means a lack of information such as a partial alignment or incomplete sequence and do not contribute to the prediction.
Please check the sequences that have been chosen. If the sequences are too diverged from your query or the alignment is questionable, we suggest you modify the fasta-formatted file above and resubmit.

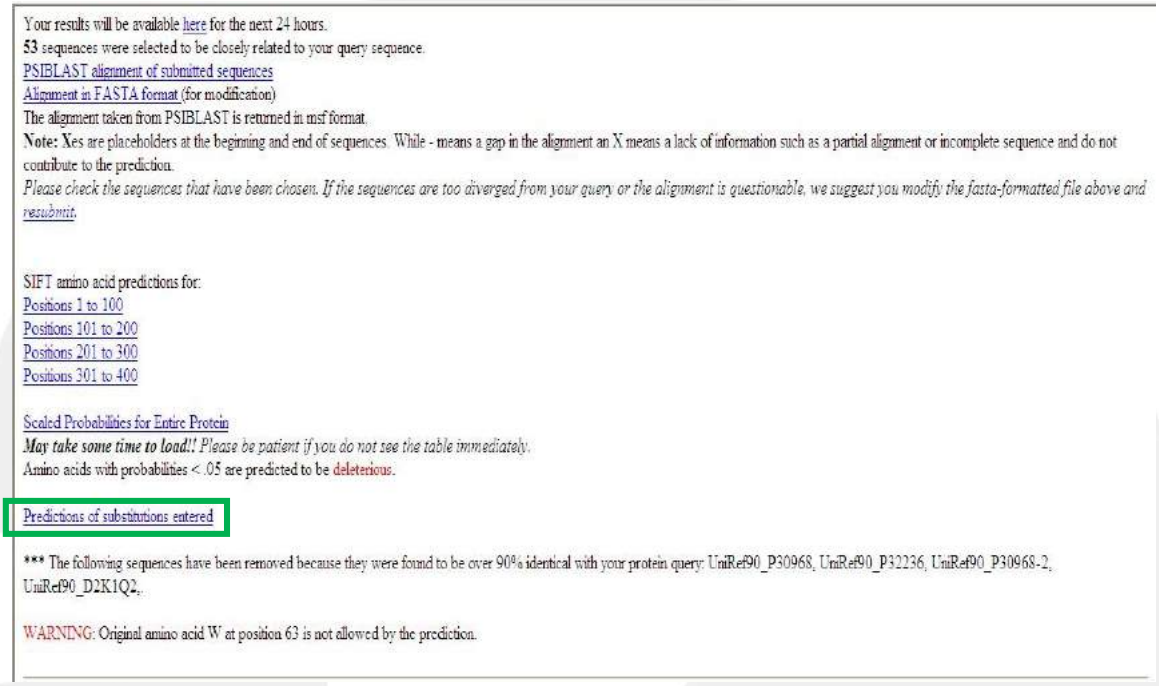
SIFT amino acid predictions for:
Positions 1 to 100
Positions 101 to 200
Positions 201 to 300
Positions 301 to 400

Scaled Probabilities for Entire Protein
***May take some time to load!!** Please be patient if you do not see the table immediately.*

http://sift.jcvi.org

Como estamos analisando uma variante específica e não as diversas alterações descritas nessa proteína podem clicar diretamente em “Predictions of substitutions entered”. Uma nova página será aberta com o resultado dessa análise de predição. Outras possibilidades de análise estão disponíveis nessa página, aconselho aos alunos clicar em diversos links e analisarem as diversas possibilidades.

FIGURA: PÁGINA COM OS DIVERSOS RESULTADOS GERADOS PELA ANÁLISE DO SIFT SEQUENCE DESTAQUE PARA A OPÇÃO PREDICTIONS OF SUBSTITUTIONS ENTERED


Your results will be available here for the next 24 hours.
53 sequences were selected to be closely related to your query sequence.
PSIBLAST alignment of submitted sequences
Alignment in FASTA format (for modification)
The alignment taken from PSIBLAST is returned in msf format.
Note: Xes are placeholders at the beginning and end of sequences. While - means a gap in the alignment an X means a lack of information such as a partial alignment or incomplete sequence and do not contribute to the prediction.
Please check the sequences that have been chosen. If the sequences are too diverged from your query or the alignment is questionable, we suggest you modify the fasta-formatted file above and resubmit.

SIFT amino acid predictions for:
Positions 1 to 100
Positions 101 to 200
Positions 201 to 300
Positions 301 to 400

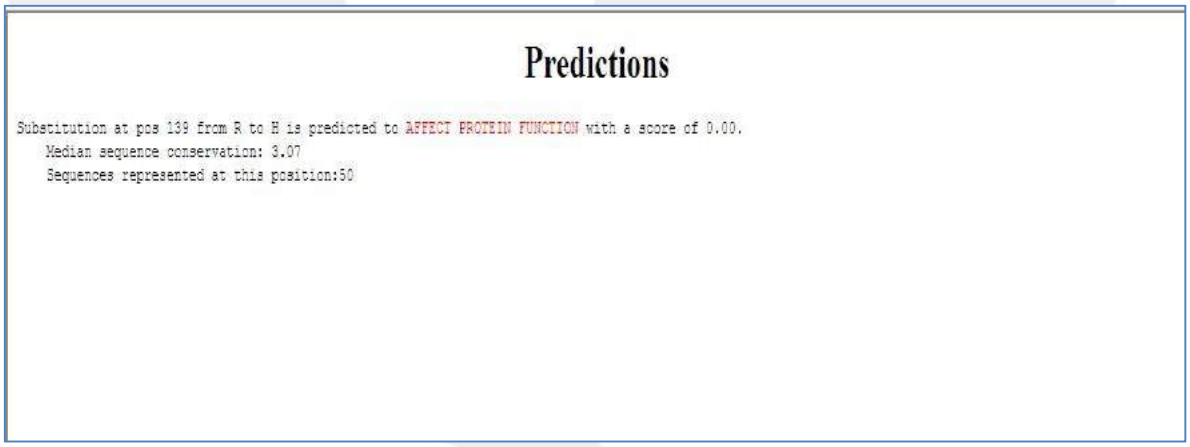
Scaled Probabilities for Entire Protein
May take some time to load!! *Please be patient if you do not see the table immediately.*
Amino acids with probabilities < .05 are predicted to be deleterious.

Predictions of substitutions entered

*** The following sequences have been removed because they were found to be over 90% identical with your protein query: UniRef90_P30968, UniRef90_P32236, UniRef90_P30968-2, UniRef90_D2K1Q2,.

WARNING: Original amino acid W at position 63 is not allowed by the prediction.

FIGURA: PÁGINA DE RESULTADOS DO PREDICTIONS OF SUBSTITUTIONS ENTERED

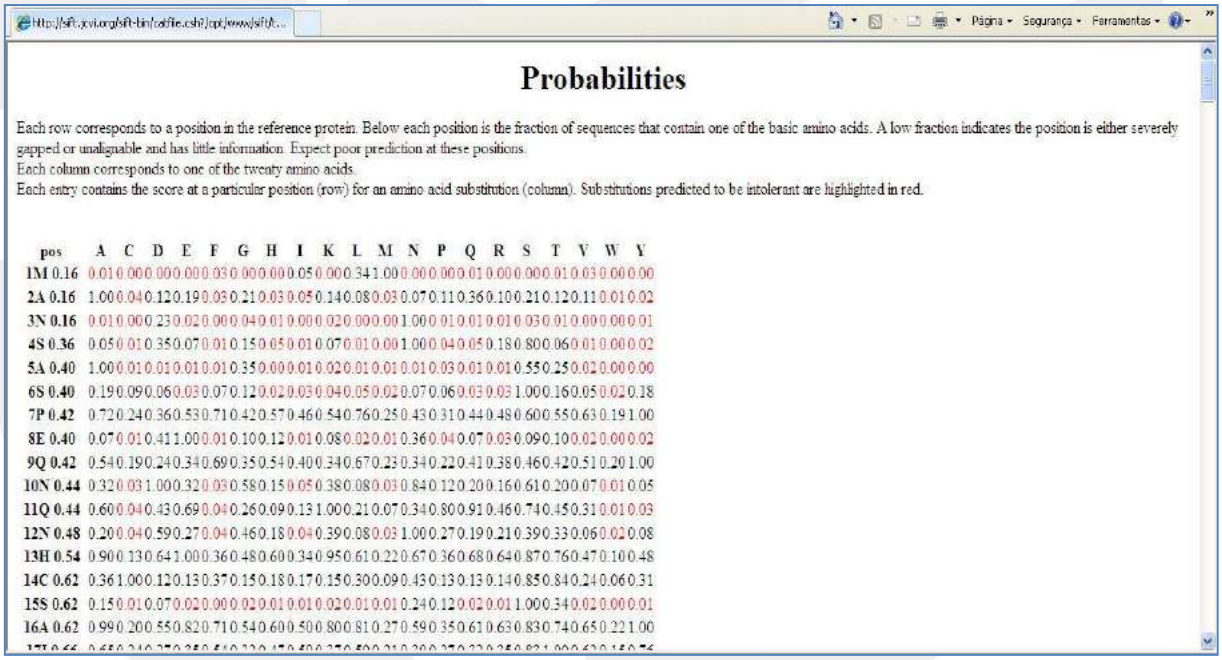


<http://sift.jcvi.org

No nosso exemplo, assim como na análise realizada pelo PolyPhen, o SIFT sugere que a variante p.R139H do GNRHR afeta a função proteica, ou seja, que essa substituição de aminoácidos é danosa para a proteína.

Apenas a título de curiosidade, se clicarmos na opção “Scaled Probabilities for Entire Protein” será aberta uma nova página com as diversas probabilidades de substituição de aminoácidos para cada posição do GNRHR. Essas probabilidades são representadas por escores, na qual quanto maior o escore (máximo 1) maior a probabilidade de a substituição de aminoácidos ser danosa para a proteína.

FIGURA: PÁGINA DE RESULTADOS DO SCALED PROBABILITIES FOR ENTIRE PROTEIN



Probabilities

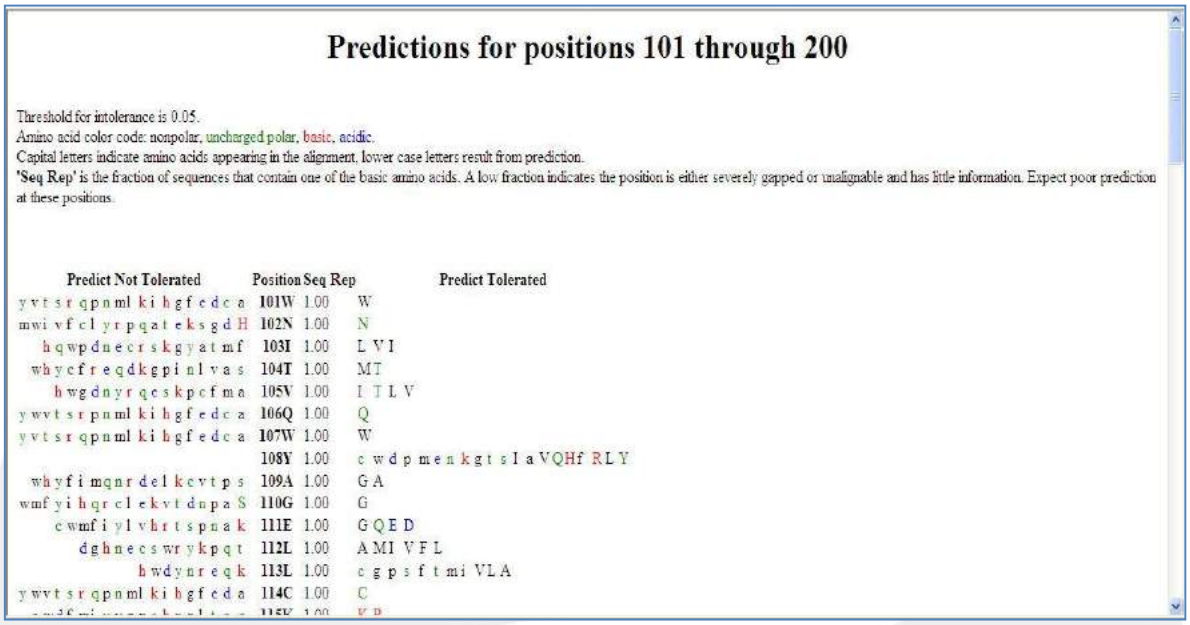
Each row corresponds to a position in the reference protein. Below each position is the fraction of sequences that contain one of the basic amino acids. A low fraction indicates the position is either severely gapped or unalignable and has little information. Expect poor prediction at these positions.
Each column corresponds to one of the twenty amino acids.
Each entry contains the score at a particular position (row) for an amino acid substitution (column). Substitutions predicted to be intolerant are highlighted in red.

pos	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1M 0.16	0.01	0.00	0.00	0.00	0.03	0.00	0.00	0.05	0.00	0.34	1.00	0.00	0.00	0.01	0.00	0.00	0.01	0.03	0.00	0.00
2A 0.16	1.00	0.04	0.12	0.19	0.03	0.21	0.03	0.05	0.14	0.08	0.03	0.07	0.11	0.36	0.10	0.21	0.12	0.11	0.01	0.02
3N 0.16	0.01	0.00	0.23	0.02	0.00	0.04	0.01	0.00	0.02	0.00	0.00	1.00	0.01	0.01	0.01	0.01	0.03	0.01	0.00	0.00	0.01
4S 0.36	0.05	0.01	0.35	0.07	0.01	0.15	0.05	0.01	0.07	0.01	0.00	1.00	0.04	0.05	0.18	0.80	0.06	0.01	0.00	0.02
5A 0.40	1.00	0.01	0.01	0.01	0.01	0.35	0.00	0.01	0.02	0.01	0.01	0.01	0.03	0.01	0.01	0.55	0.25	0.02	0.00	0.00
6S 0.40	0.19	0.09	0.06	0.03	0.07	0.12	0.02	0.03	0.04	0.05	0.02	0.07	0.06	0.03	0.03	1.00	0.16	0.05	0.02	0.18
7P 0.42	0.72	0.24	0.36	0.53	0.71	0.42	0.57	0.46	0.54	0.76	0.25	0.43	0.31	0.44	0.48	0.60	0.55	0.63	0.19	1.00
8E 0.40	0.07	0.01	0.41	1.00	0.01	0.10	0.12	0.01	0.08	0.02	0.01	0.36	0.04	0.07	0.03	0.09	0.10	0.02	0.00	0.02
9Q 0.42	0.54	0.19	0.24	0.34	0.69	0.35	0.54	0.40	0.34	0.67	0.23	0.34	0.22	0.41	0.38	0.46	0.42	0.51	0.20	1.00
10N 0.44	0.32	0.03	1.00	0.32	0.03	0.58	0.15	0.05	0.38	0.08	0.03	0.84	0.12	0.20	0.16	0.61	0.20	0.07	0.01	0.05
11Q 0.44	0.60	0.04	0.43	0.69	0.04	0.26	0.09	0.13	1.00	0.21	0.07	0.34	0.80	0.91	0.46	0.74	0.45	0.31	0.01	0.03
12N 0.48	0.20	0.04	0.59	0.27	0.04	0.46	0.18	0.04	0.39	0.08	0.03	1.00	0.27	0.19	0.21	0.39	0.33	0.06	0.02	0.08
13H 0.54	0.90	0.13	0.64	1.00	0.36	0.48	0.60	0.34	0.95	0.61	0.22	0.67	0.36	0.68	0.64	0.87	0.76	0.47	0.10	0.48
14C 0.62	0.36	1.00	0.12	0.13	0.37	0.15	0.18	0.17	0.15	0.30	0.09	0.43	0.13	0.13	0.14	0.85	0.84	0.24	0.06	0.31
15S 0.62	0.15	0.01	0.07	0.02	0.00	0.02	0.01	0.01	0.02	0.01	0.01	0.24	0.12	0.02	0.01	1.00	0.34	0.02	0.00	0.01
16A 0.62	0.99	0.20	0.55	0.82	0.71	0.54	0.60	0.50	0.80	0.81	0.27	0.59	0.35	0.61	0.63	0.83	0.74	0.65	0.22	1.00

<http://sift.jcvi.org

Outro link interessante é o “SIFT amino acid predictions” encontraremos também as diversas possibilidades de substituição de aminoácidos, porém, ao invés de serem dados escores, as substituições são classificadas como possivelmente toleradas e não toleradas.

FIGURA: PÁGINA DE RESULTADOS DO SIFT AMINO ACID PREDICTIONS



Predictions for positions 101 through 200

Threshold for intolerance is 0.05.
Amino acid color code: nonpolar, uncharged polar, basic, acidic.
Capital letters indicate amino acids appearing in the alignment, lower case letters result from prediction.
'Seq Rep' is the fraction of sequences that contain one of the basic amino acids. A low fraction indicates the position is either severely gapped or unalignable and has little information. Expect poor prediction at these positions.

Predict Not Tolerated	Position	Seq Rep	Predict Tolerated
y v t s r q p n m l k i h g f e d c a	101W	1.00	W
m w i v f c l y r p q a t e k s g d H	102N	1.00	N
h q w p d n e c r s k g y a t m f	103I	1.00	L V I
w h y c f r e q d k g p i n l v a s	104T	1.00	M T
h w g d n y r q c s k p c f m a	105V	1.00	I T L V
y w v t s r p n m l k i h g f e d c a	106Q	1.00	Q
y v t s r q p n m l k i h g f e d c a	107W	1.00	W
	108Y	1.00	c w d p m e n k g t s I a V Q H f R L Y
w h y f i m q n r d e l k c v t p s	109A	1.00	G A
w m f y i h q r c l e k v t d n p a S	110G	1.00	G
c w m f i y l v h r t s p n a k	111E	1.00	G Q E D
d g h n e c s w r y k p q t	112L	1.00	A M I V F L
h w d y n r e q k	113L	1.00	c g p s f t m i V L A
y w v t s r q p n m l k i h g f e d a	114C	1.00	C

<http://sift.jcvi.org/sift-bin/catfile.csh?/opt/www/sift/tmp/14714.aatable2>.

Diversos outros programas estão disponíveis para análise de substituição de aminoácidos. Aconselho os alunos a entrarem em diversos programas existentes e realizarem testes nos mesmos, a fim de entender como funcionam e desenvolverem uma postura crítica para a interpretação dos resultados.

REFERÊNCIAS

ANDALIA, R. C.; JORGE, R. A. Bioinformática: em busca de los secretos moleculares de La vida. Acimed, v. 12, n. 6. Disponível em:<http://bvs.sld.cu/revistas/aci/vol12_6_04/aci02604.htm>. Acesso em: 10 nov. 2011.

ANDRADE, G. G. L. A importância da bioinformática no setor farmacêutico. Disponível em:<http://www.webartigos.com/artigos/a-importancia-da-bioinformatica-no-setor- farmaceutico/8172/>. Acesso em: 10 nov. 2011.

ANDRADE, J. M. S. Código Genético e Síntese Proteica. Disponível em:<http://genetica.ufcspa.edu.br/indexmed.html>. Acesso em: 05 out. 2011.

BARALLE, D.; LUCASSEN A.; BURATTI, E. Missed threads. The impact of pre-mRNA splicing defects on clinical practice, EMBO Rep, v. 10, n. 8, p. 810-816, 2009.

BAXEVANIS, A. D.; OUELLETE, B. F. F. Bioinformatics: A practical guide to the analysis og genes and proteins. 2.. ed., Wiley-interscience, 2001, 470 p.BAXEVANIS, A. D.; OUELLETTE, B. F. F. Bioinformatics: A practical guide to the analysis of genes and proteins. 3. ed. Wiley-insterscience, 2005, 540 p.

BIBGEN. Introducción a la bioinformática. Disponível em: <http://bvs.isciii.es/bib-gen/actividades/curso_virtual/introduction/biinformatica.htm>. Acesso em: 05 nov. 2011.

BIOMEDICINA. O dogma central da biologia molecular. Disponível em:<http://www.biomedicina3l.com/2010/11/o-dogma-central-da-biologia-molecular.html>. Acesso em: 10 out. 2011.

BLAST PROGRAM SELECTION GUIDE. Disponível :<http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Prog SelectionGuid>. Acesso em: 10 dez. 2011.

BRIAN MCKNIGHT BRAIN INSTITUTE. Conceitos Moleculares do Material Genético e Farmacodinâmica Médica. Disponível em:<http://www.sistemanervoso.com/pagina.php?secao=5&materia_id=95&materiaver=1>. Acesso em: 10 out. 2011.0

famart
GRUPO EDUCACIONAL