http://i.imgur.com/2BLvvHj.png

**Universidade do Minho**

**Escola de Engenharia – Departamento de Informática**

**Mestrado em Bioinformática**

**Laboratório de Bioinformática/Algoritmos para análise de sequências Biológicas**

Grupo 6 composto por:

Daniel Varzim PG27662

João Silva PG27661

Catarina Lemos PG27664

Este trabalho tem como objetivo a anotação funcional do genoma do organismo de interesse, ou seja, a atribuição de funções ao máximo número de genes possível e proteínas do organismo.

A zona do genoma que iremos estudar é a 1162601 até 1422700, locus\_tag NGO1213 até NGO1455.

Visto isto iremos começar com uma breve introdução ao organismo em questão.

**Neisseria gonorrhoeae**

A Neisseria gonorrhoeae é uma bactéria Gram-negativa do tipo diplococos e um patogéneo que tem como hospedeiro obrigatório o Homem. Inicialmente, coloniza e invade o epitélio da mucosa reprodutora e causa uma das doenças sexualmente transmissíveis mais comuns, a gonorreia. Em cada ano, cerca de 6 milhões de casos de gonorreia são reportados mundialmente, sendo estimada a segunda infeção bacteriana e doença sexualmente transmissível mais prevalente no mundo. Até agora, ainda não existe uma vacina disponível, considerando-se a sua transmissão e infeção um problema global de saúde pública.

A prevalência da infeção por N. gonorrhoeae pode variar de acordo com o gênero, faixa etária, etnia, presença de sintomas, local do estudo e o teste empregado para o diagnóstico, o que dificulta sua real determinação. Os indivíduos mais afetados pela infeção gonocócica são as mulheres com idade igual ou inferior a 25 anos [4].

Em cada ano, cerca de 6 milhões de casos de gonorreia são reportados mundialmente, sendo estimada a segunda infeção bacteriana e doença sexualmente transmissível mais prevalente o mundo [[2](#_ENREF_2)].

O diagnóstico de infeção é feito, geralmente, através de uma cultura bacteriana ou ensaios de amplificação de ácidos nucleicos (NAAT). Estes testes são realizados em ambientes controlados de laboratório, com resultados geralmente disponíveis em poucos dias. Na maioria dos casos os médicos pedem então que o paciente compareça para o tratamento ser iniciado. Em clínicas de zonas urbanas a maioria dos pacientes com diagnóstico de STI’s recebe tratamento atempadamente [1].

Até agora, ainda não existe uma vacina disponível, considerando-se a sua transmissão e infeção um problema global de saúde pública [3].

Devido ao facto da infeção gonocócica em mulheres ser frequentemente assintomática, um componente importante para o controlo da doença é o seu rastreamento de mulheres com risco aumentado para DST. O rastreamento em homens ainda é muito discutido. No entanto, nos últimos anos, tem sido preconizado em determinadas populações de jovens do sexo masculino, como por exemplo, entre aqueles que se apresentam para o serviço militar [5].

De modo a tornar a análise e recolha de dados sobre a nossa porção do genoma menos demorada e cansativa o grupo desenvolveu e utilizou os seguintes scripts:

**Scripts criados**

No decorrer do desenvolvimento do trabalho o grupo foi desenvolvendo vários scripts de modo a auxiliar a pesquisa e documentação de diversos atributos dos vários genes e proteínas.

Inicialmente precisávamos de criar um script que recolhesse ficheiros genbank do NCBI de todos os genes da nossa porção do genoma. Devido ao facto do grupo 5 já ter este script criado decidimos utiliza-lo (“Fetch\_gb\_records”), criando assim uma pasta com todos os ficheiros genbank referentes a todos os genes da nossa secção, pelo que o nome de cada ficheiro é o Locus\_Tag do gene a que se refere.

Utilizando também outro script do grupo 5 (“Validate”) fomos capazes de comparar a tabela “ProteinTable864\_169534” fornecida pelo docente com os dados dos ficheiros genbank gerados, obtendo um ficheiro “log\_gb records”. Este script faz a validação de alguns dados da tabela, uma vez que os compara com os dados dos ficheiros genbank. Com isto fomos capazes de verificar que o gene com locus\_tag NGO1285 apresentava uma localização no genoma diferente da informação recolhida no seu ficheiro genbank. Também o gene com locus\_tag NGO1442 apresentava informação diferente no que diz respeito à sua localização no genoma.

Em seguida criamos um script chamado “pubmed”, que é capaz de efectuar as seguintes funções:

-NumeroArtigos(nome): Função que faz a procura no PubMed de todos os artigos que contêm informações sobre o “nome” que escolhemos, efectuando um print do número de artigos existentes.

-DownloadArtigosIDs (nome, numArtigos): Função que retira todos os Ids, colocando-os numa lista, de todos os artigos com aquele “nome” até chegar ao número de artigos escolhido em “numArtigos”.

-ObterRegistos (idlist): Função que recebe o “idlist” gerado na função “DownloadArtigosIDs”, e vai busar ao Medline todos os records com o id presente em “idlist”.

-VerRegistos (records): Função que vai imprimir alguns dados sobre os records gerados na função “ObterRegistos”, como titulo, autores e source do artigo.

-ProcuraTitulo (records): Função que procura em todos os records gerados, um titulo ou palavra, que pode ser introduzido pelo utilizador, efetuando o print do record[“SO”] do artigo se a palavra estiver no titulo.

-ProcuraAutor (records): Função que procura em todos os records gerados, um autor (s), que pode ser introduzido pelo utilizador, efetuando o print do record[“SO”] do artigo no caso do autor estar presente.

-Abstract (Titulo): Função que acede ao pubmed e procura o artigo com o “Titulo”. Se encontrado, acede ao MedLine e faz print ao abstract do artigo.

O grupo criou um script simples para sabermos um pouco mais sobre a taxonomia do organismo e da sua linhagem.

-lineage (Organism): Função que acede à base de dados Taxonomy e retorna a linhagem do organismo.

Um script bastante importante para o desenvolvimento deste trabalho foi o “GeneInfo\_Genbank” que recolhe informação importante sobre os genes e proteínas e possui as seguintes funções:

-Get\_Ids (): Função que vai à pasta que contem todos os ficheiros genbank sobre os genes, e vai aceder a cada um dos ficheiros genbank, colocando numa lista os Ids de todos os genes.

-Get\_Protein\_Names ():Função que vai à pasta que contem todos os ficheiros genbank sobre os genes, e vai aceder a cada um dos ficheiros genbank, colocando numa lista os nomes das proteínas associadas a cada gene.

-Get\_Gene\_lenght ():Função que vai à pasta que contem todos os ficheiros genbank sobre os genes, e vai aceder a cada um dos ficheiros genbank, colocando numa lista o tamanho de todos os genes.

-Get\_Gene\_Location ():Função que vai à pasta que contem todos os ficheiros genbank sobre os genes, e vai aceder a cada um dos ficheiros genbank, colocando numa lista a localização e strand de todos os genes.

-Get\_CDD ():Função que vai à pasta que contem todos os ficheiros genbank sobre os genes, e vai aceder a cada um dos ficheiros genbank, colocando numa matriz, todos os CDD associados a cada gene.

-Este script também contém várias funções que gravam em diferentes ficheiros os dados gerados, de modo a facilitar o preenchimento da tabela final.

Um script bastante importante para o desenvolvimento deste trabalho foi o “Uniprot\_Info” (Utilizando algumas funções retiradas do grupo1) que recolhe informação importante sobre os genes e proteínas e possui as seguintes funções:

-Get\_Gene\_IDs ():Função que vai à pasta que contem todos os ficheiros genbank sobre os genes, e vai aceder a cada um dos ficheiros genbank, colocando numa lista os Ids de todos os genes.

-uniprot\_ID (): Função que acede à uniprot utilizando os ids gerados na função “Get\_Gene\_IDs”, e vai retornar uma lista com todos os ids da uniprot referentes a cada id do gene.

-infor\_uniprot (): Função que acede ao ficheiro .txt retirado da uniprot com informação sobre todo o proteoma da Neisseria gonorrhoeae e através de um módulo “Uniprot\_Parser” retirado do grupo1 vai efetuar o parse de todos os records cujo id é referente à nossa porção do genoma, retirando alguma informação usada para completar a tabela final.

-more\_info\_uniprot ():Função que acede ao ficheiro .xml retirado da uniprot com informação sobre todo o proteoma da Neisseria, retirando alguma informação de todos os records cujo id é referente à nossa porção do genoma, usando essa informação para completar a tabela final.

De salientar que pedimos ao grupo 1 para nos transformar a informação retirada das funções “info\_uniprot” e “more\_info\_uniprot” em ficheiros excel, sem a ajuda deles a análise da informação seria muito mais complicada.

Por último criámos dois scripts para nos auxiliar na análise do blast dos genes cuja função era desconhecida.

Primeiro script de nome “Blast” possui as seguintes funções:

-blast (filename): Função que efetua o blastp de um ficheiro genbank à escolha pelo utilizador.

-get\_locusTag (record) e blastp (genes): Funções retiradas do grupo5 que fazem o blastp de todos os ficheiros os genes associados à nossa secção do genoma.

Segundo script de nome “Parse\_Blast” só contem uma função:

-parsing\_blast (filename): Função que efetua a análise de todos os blast records gerados com e.value inferior a 0.05 de um ficheiro gerado pelo script anterior. Imprime informações como: alignment.title, alignment.length, hsp.expect, hsp.query, hsp.match, hsp.sbjct.

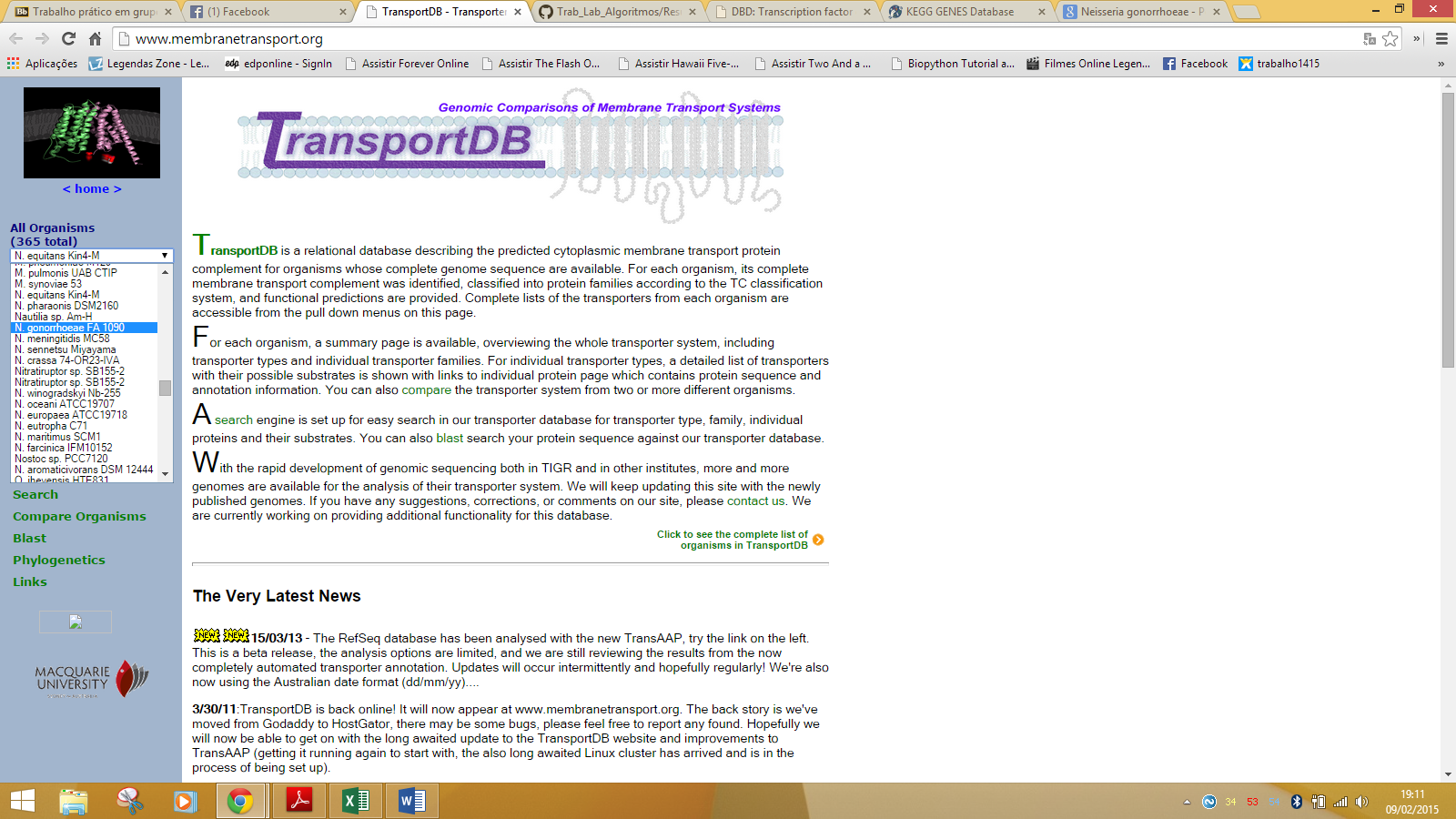
Após a obtenção de toda esta informação procedemos à confirmação/atribuição da função de cada gene/proteína. Em seguida iremos demonstrar sucintamente como chegamos à função dos genes/proteínas referentes à nossa porção do genoma.

**Funções das proteínas**

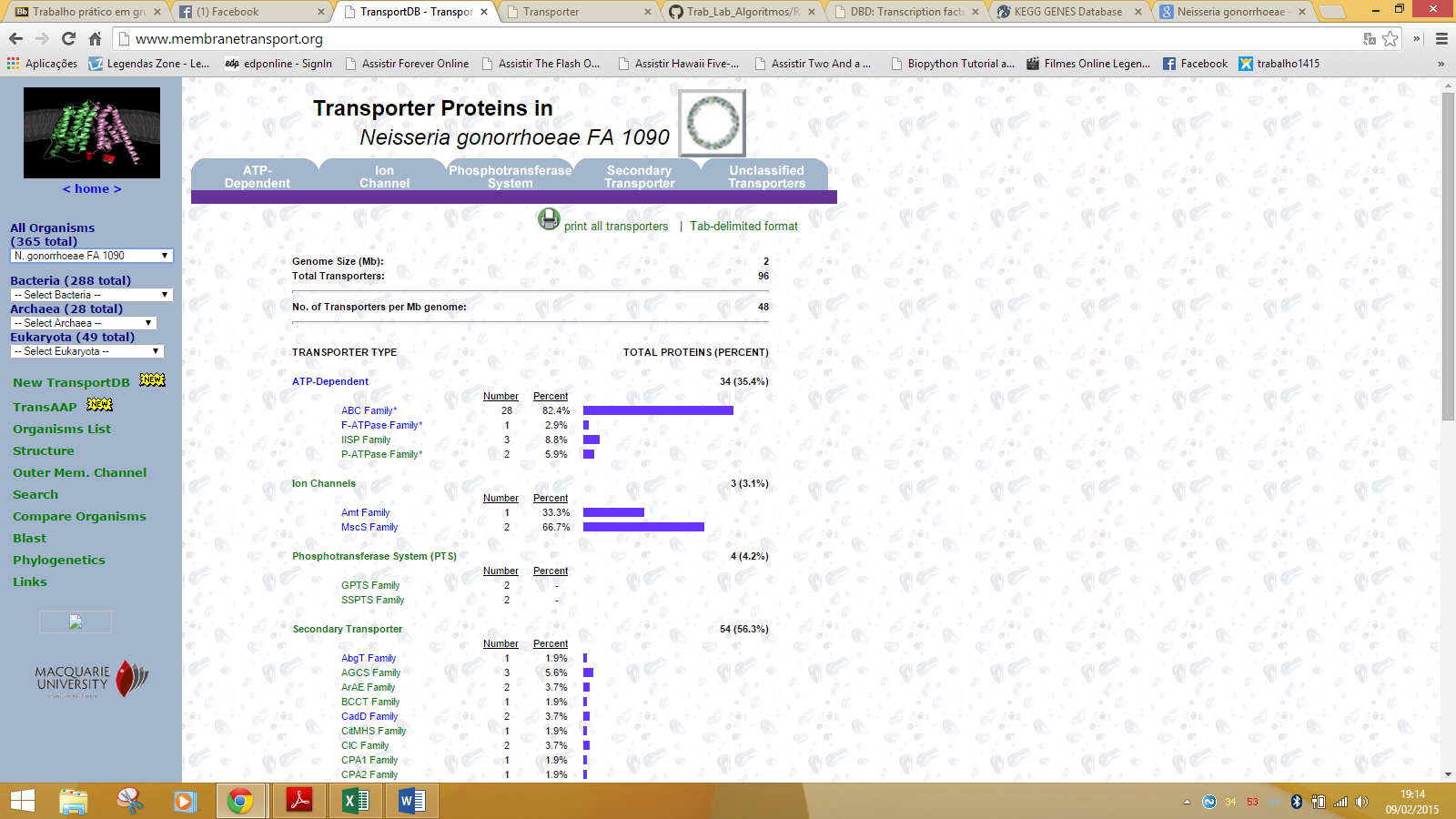
Podemos dividir estas funções em várias categorias e existem várias maneiras de as descobrir. O grupo decidiu em primeiro lugar ir a bases de dados específicas como vamos ver a seguir (Parte I). Depois resolvemos inferir as funções de acordo com a família de cada uma das proteínas devido a dificuldade de encontrar a função em si (Parte II). Depois destas duas partes feitas, verificamos que ainda não sabíamos e não tinhas certeza da função de algumas proteínas, portanto fomos fazer um BLAST a essas mesmas (Parte III).

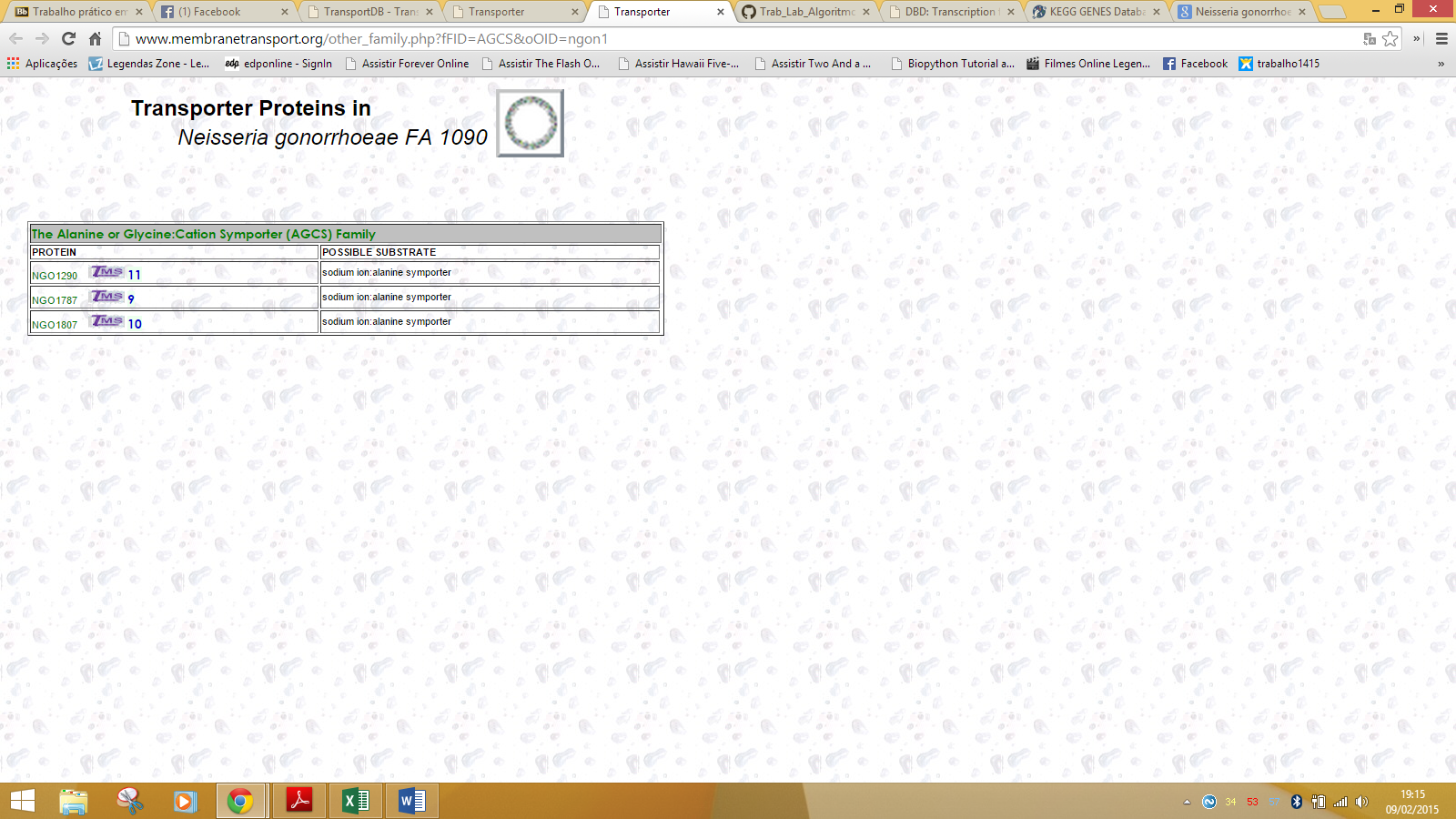
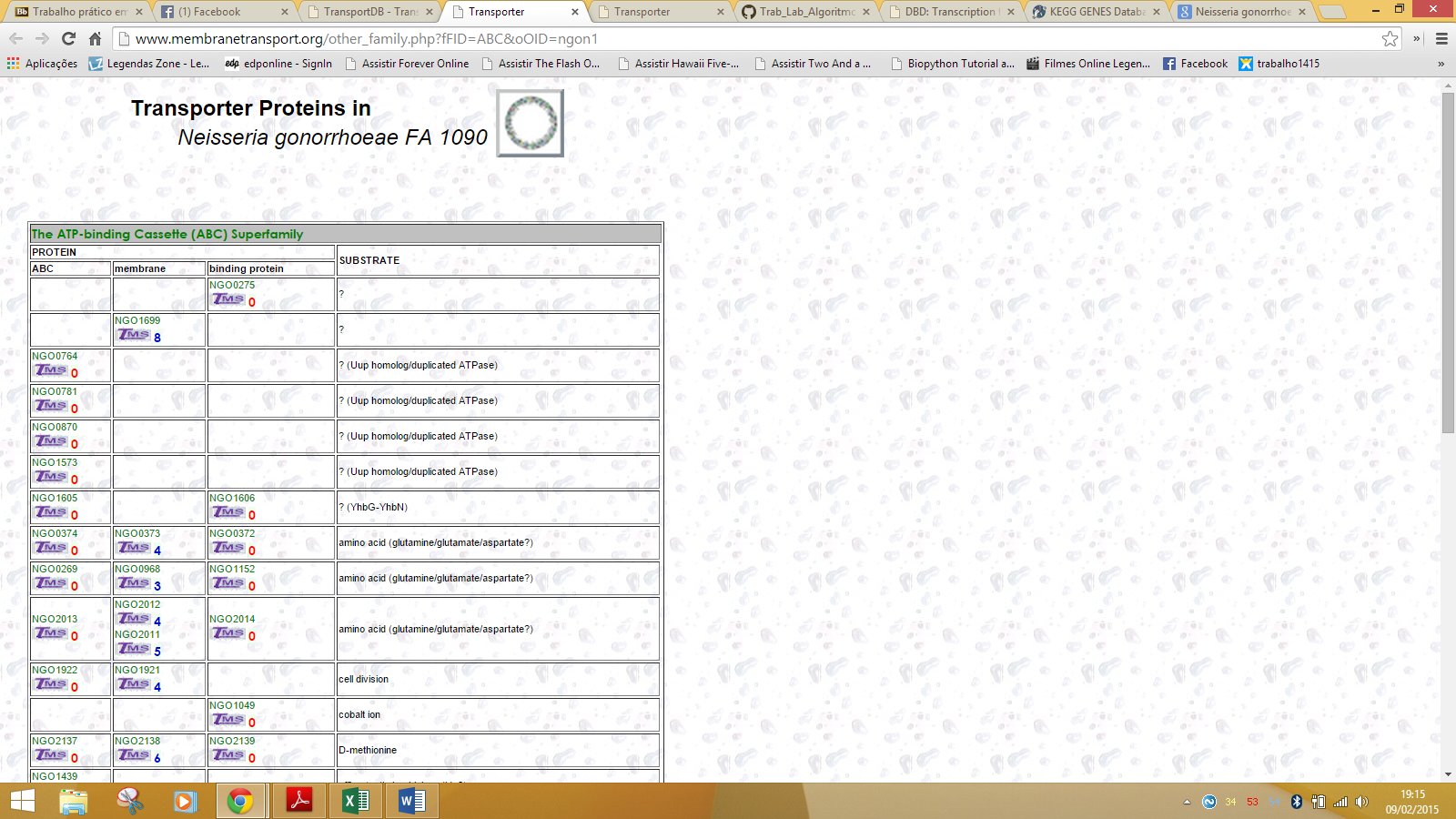
Parte I:

* Proteínas com **função transportadora** (<http://www.membranetransport.org/>), procurando depois em bactérias por Neisseria gonorrhoeae

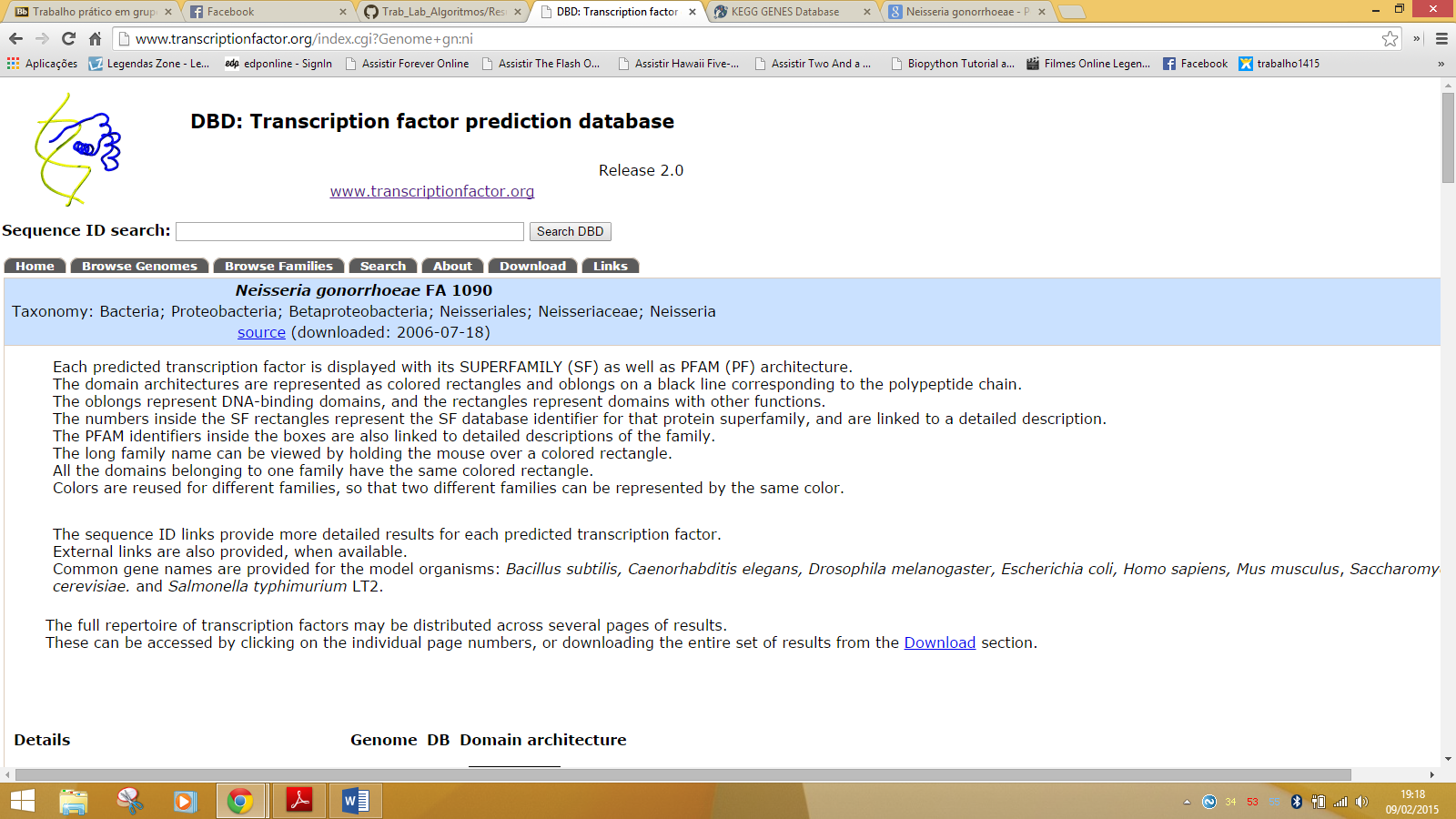


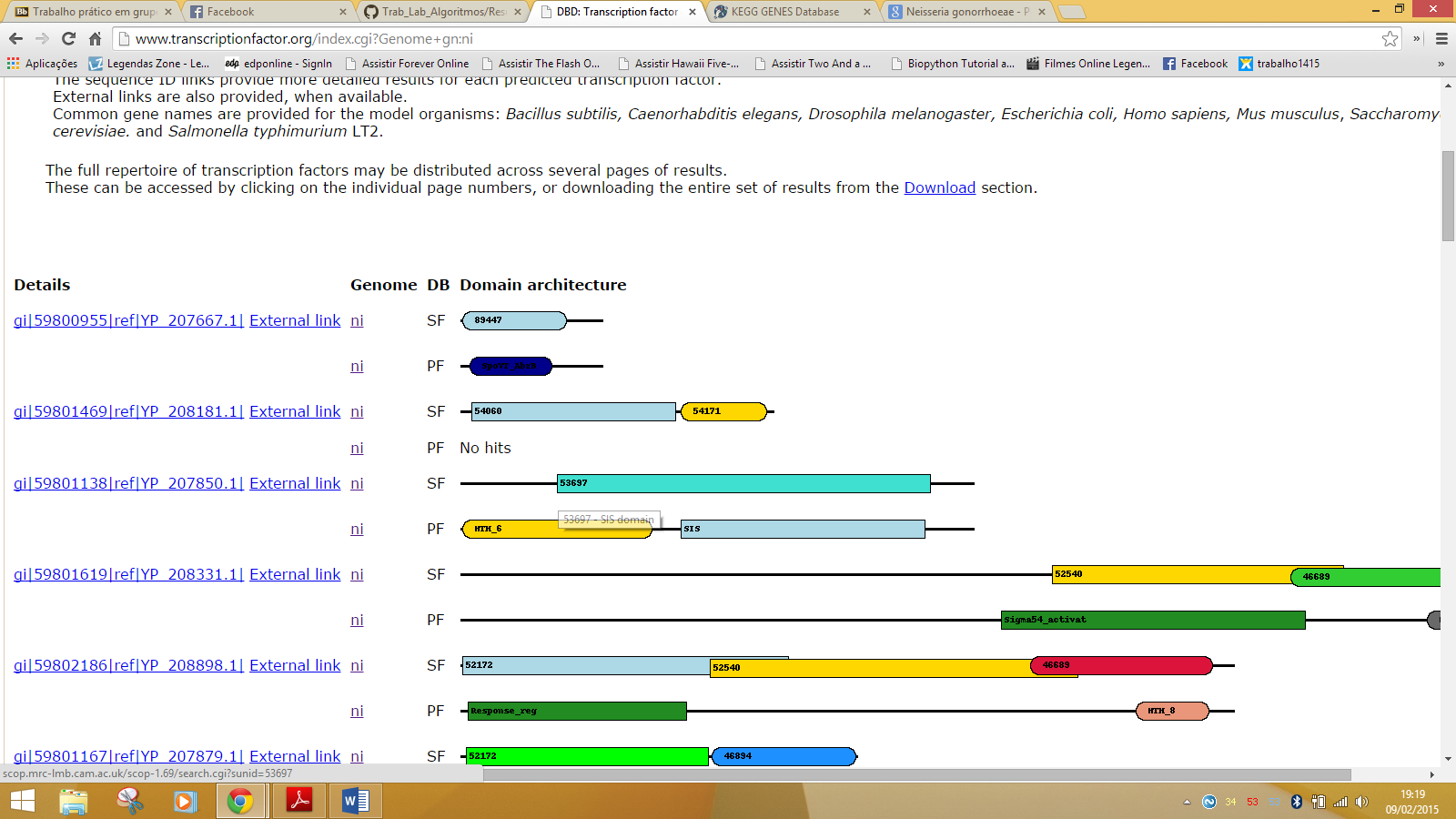
Depois deste passo, encontraremos a seguinte página:



Ao procuramos por cada um dos “Transporter Type”, vamos encontrar que locus\_tag se tem este tipo de função. Como por exemplo:

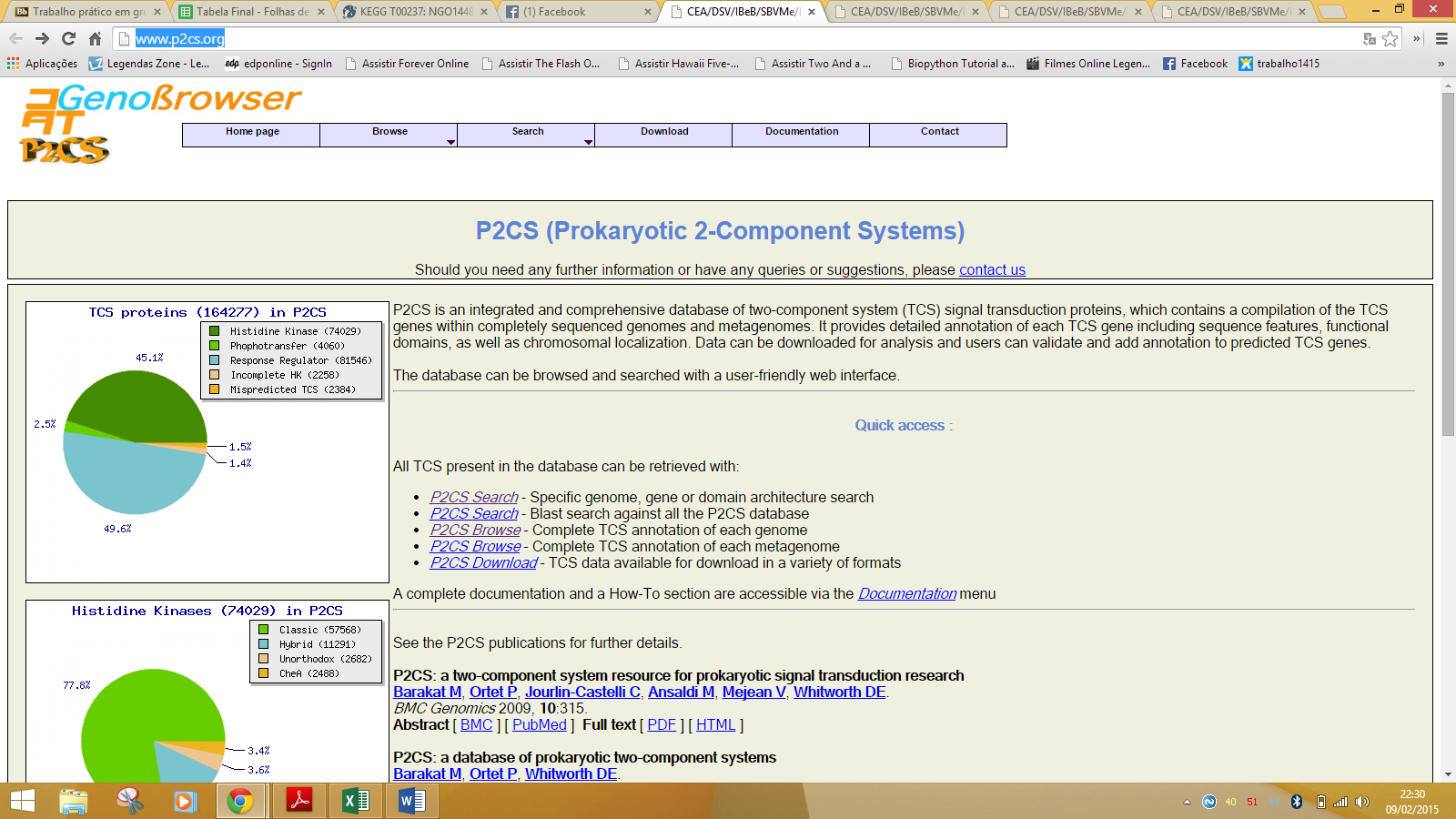
* Proteínas com **função de transcrição** (<http://www.transcriptionfactor.org/index.cgi?Browse>) e procurando por Neisseria gonorrhoeae vamos ter à seguinte página:

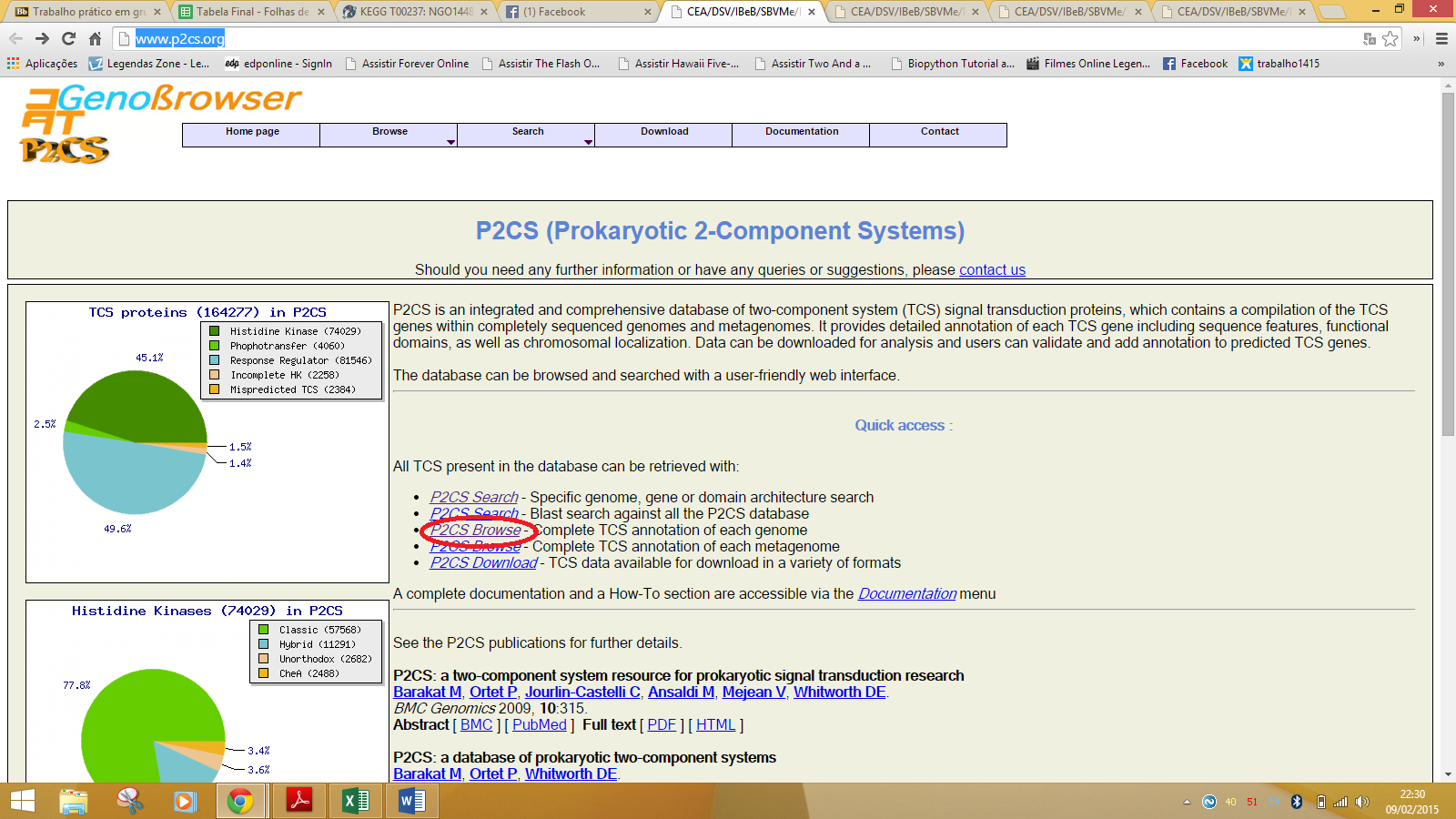




Após este passo, procuramos através do gi de cada das proteínas e caso este seja encontrado, significa que essa mesma proteína tem função de sinalização.

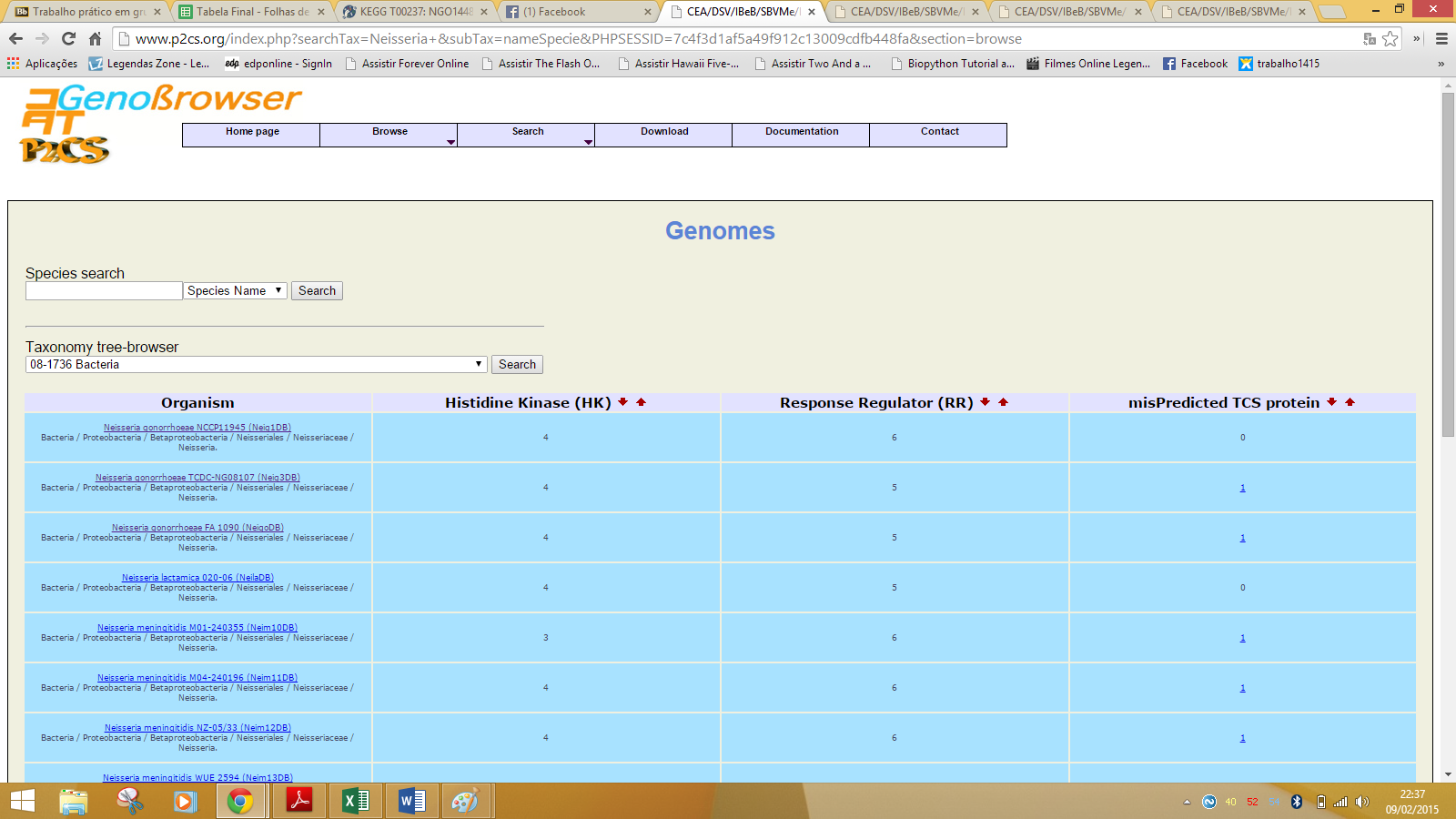
* Proteinas com **função de sinalização** (<http://www.p2cs.org/>)





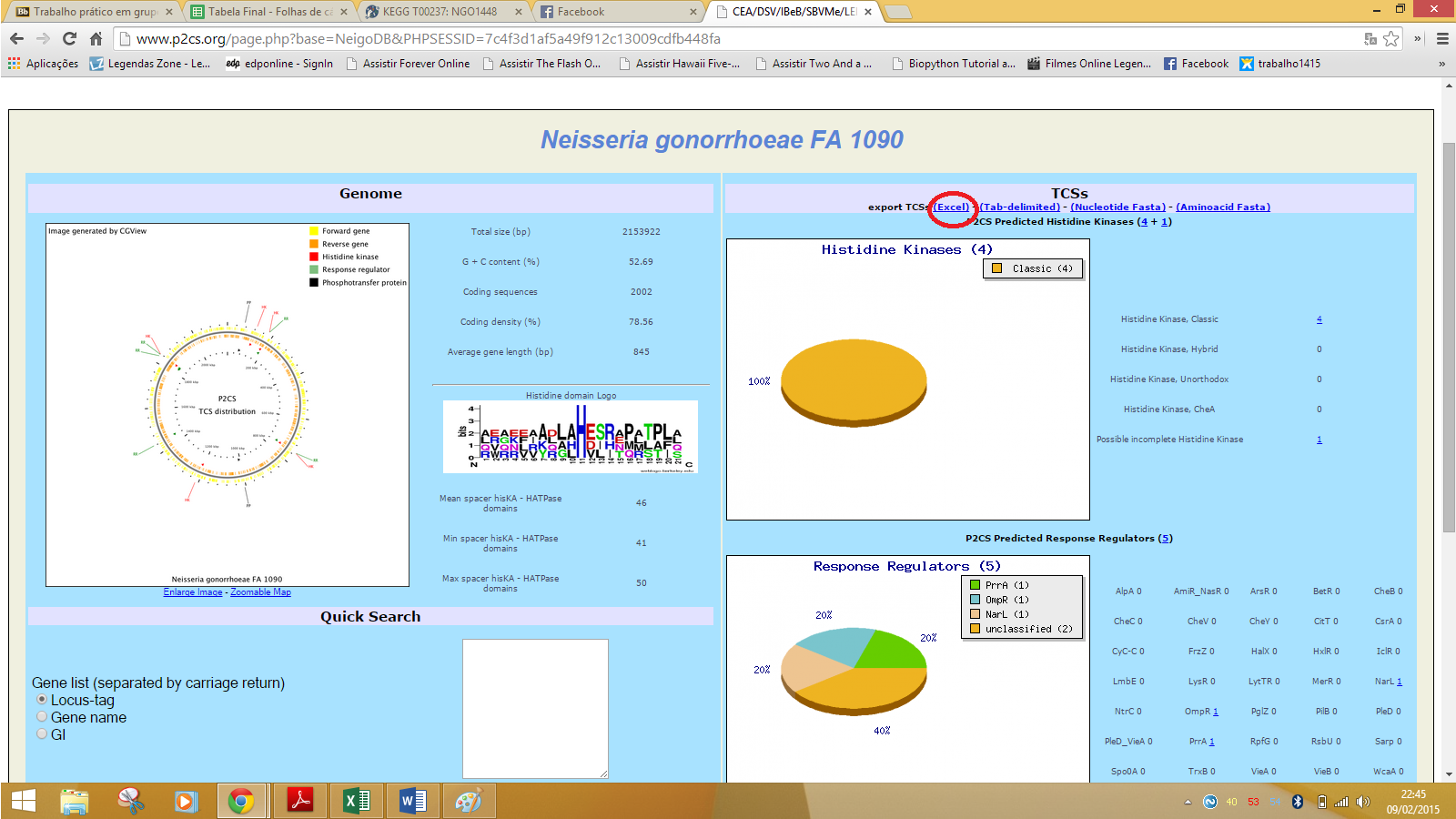
Depois de entramos no link acima referido, vamos clicar em P2CS Browse.

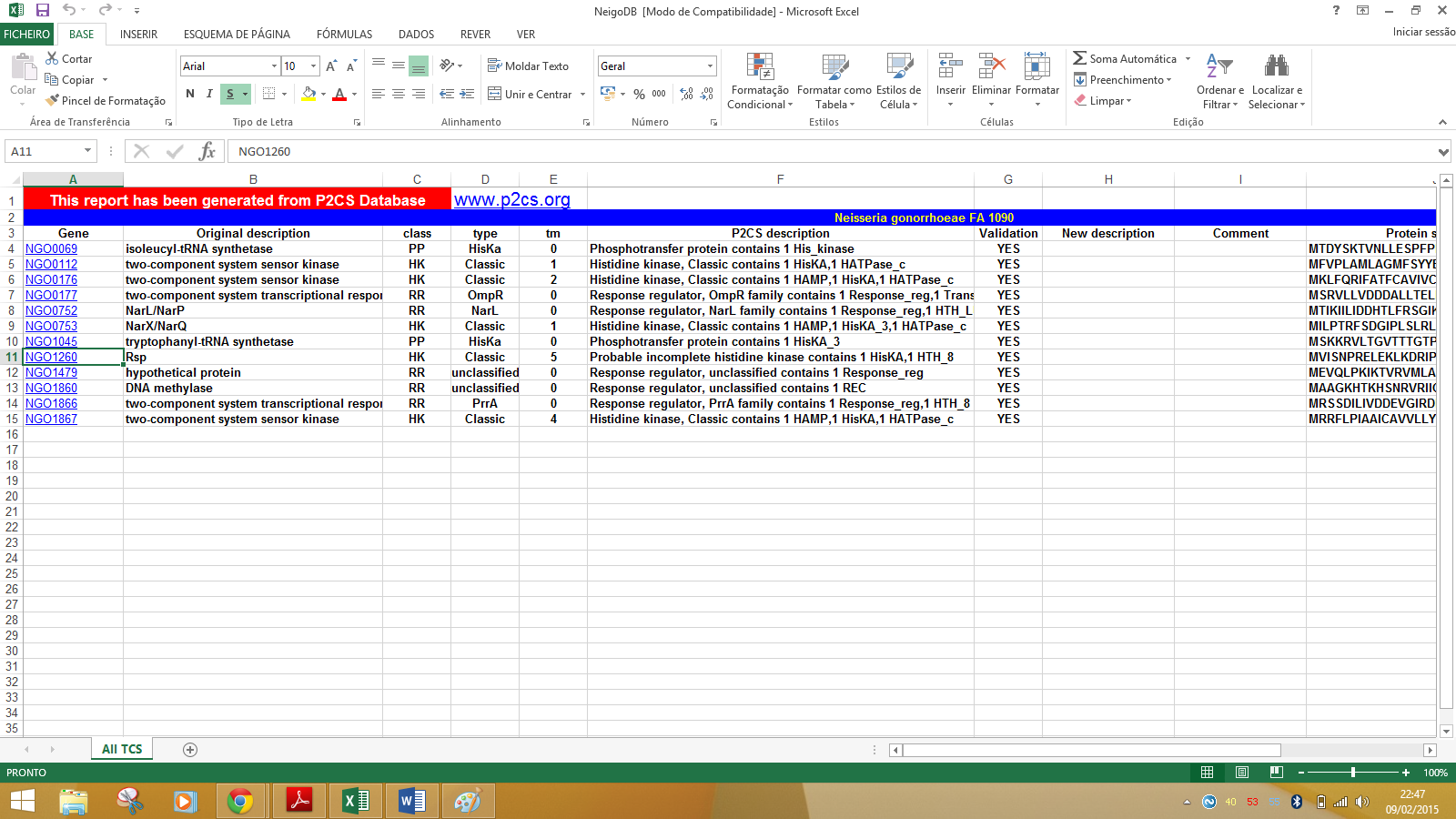
Com o passo anterior efetuado vai-nos aparece a página da imagem seguinte. Procuramos pelo organismo que pretendemos pesquisar.

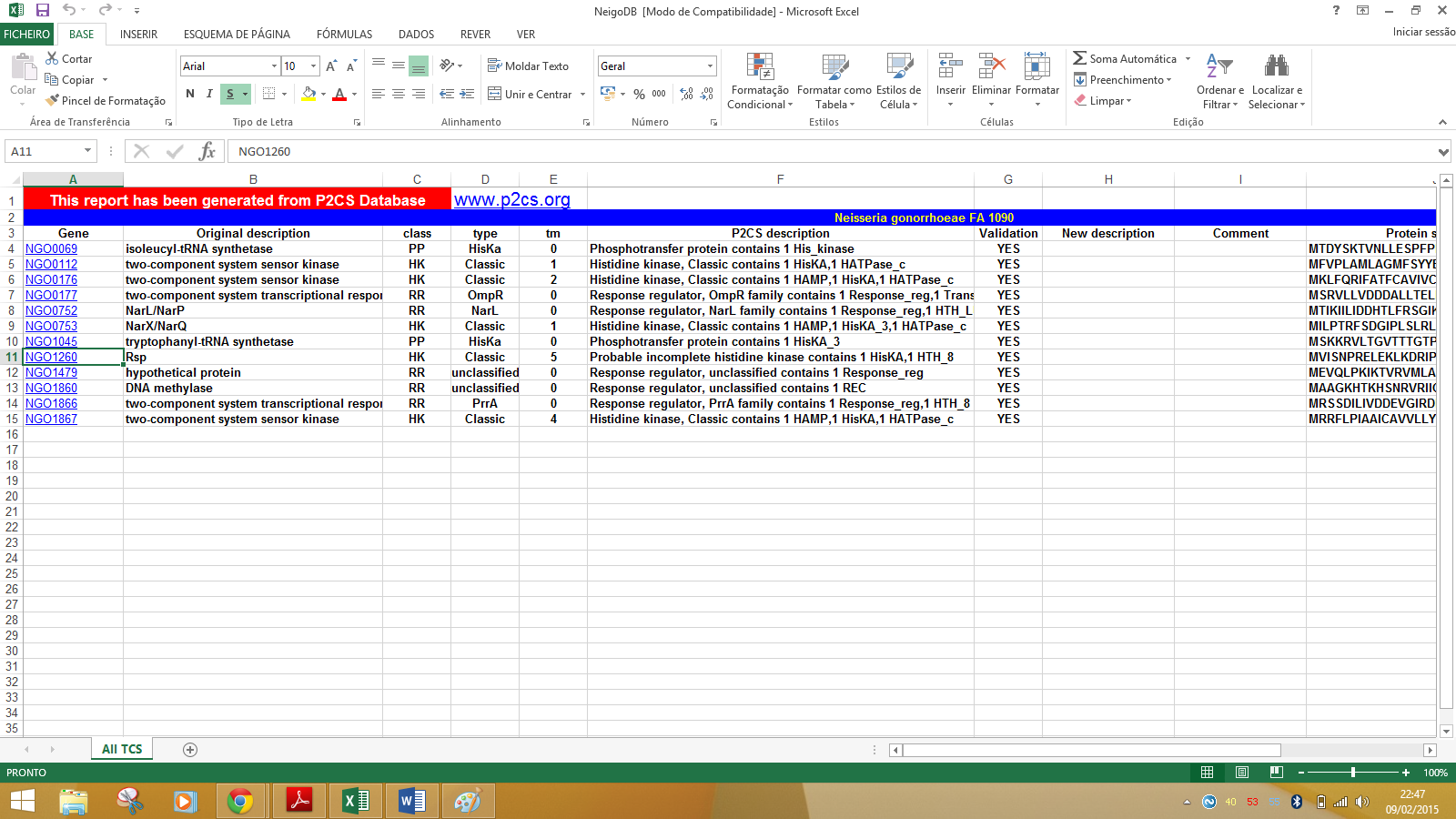


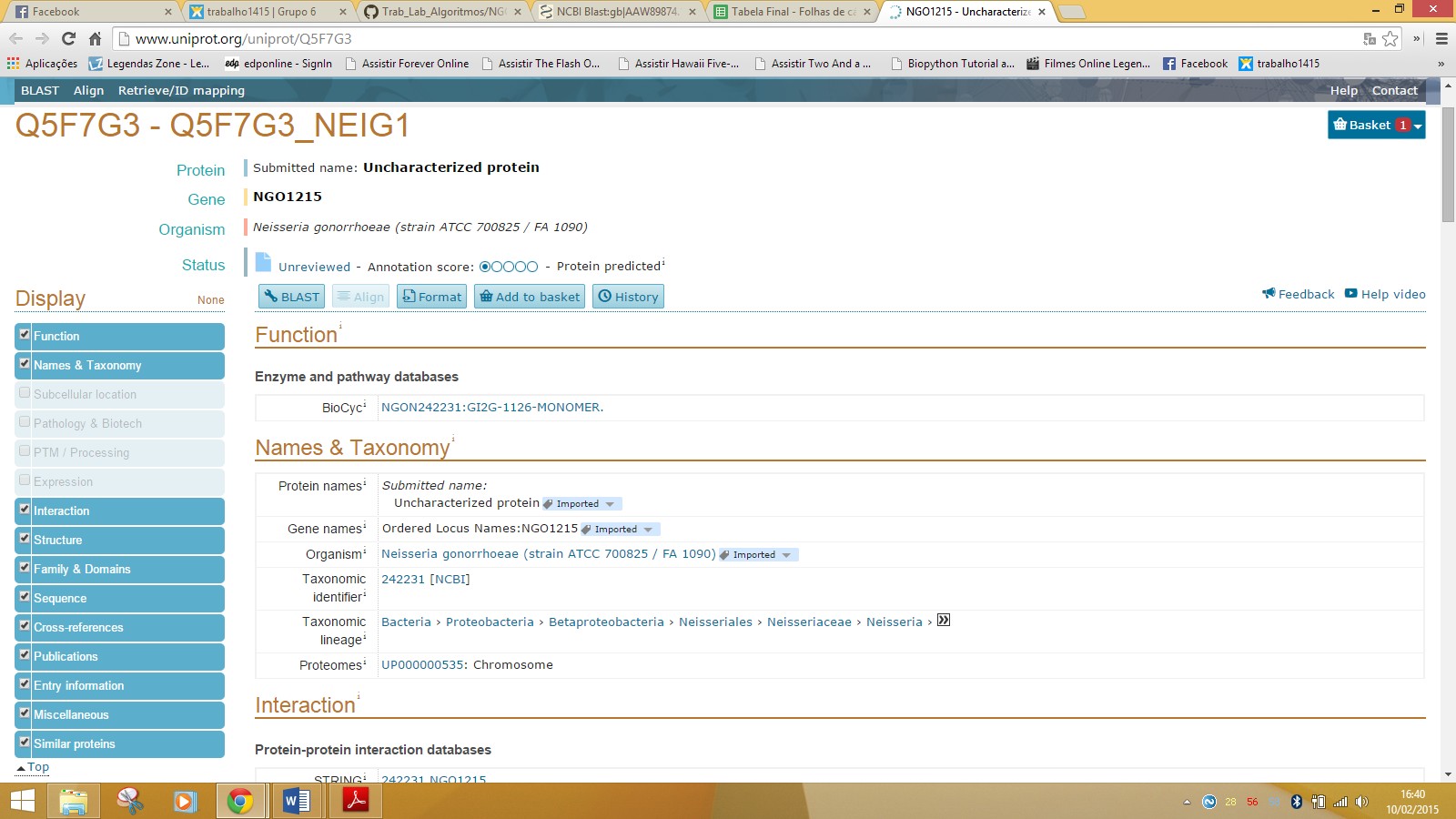
Selecionamos o nome da espécie que queremos, “[Neisseria gonorrhoeae FA 1090 (NeigoDB)](http://www.p2cs.org/page.php?base=NeigoDB&PHPSESSID=7c4f3d1af5a49f912c13009cdfb448fa)”, neste caso. Depois disto vamos ter à seguinte página.

Já nesta página vamos ter a opção de extrair um documento Excel, onde teremos os locus\_tag referentes a esta função de sinalização.

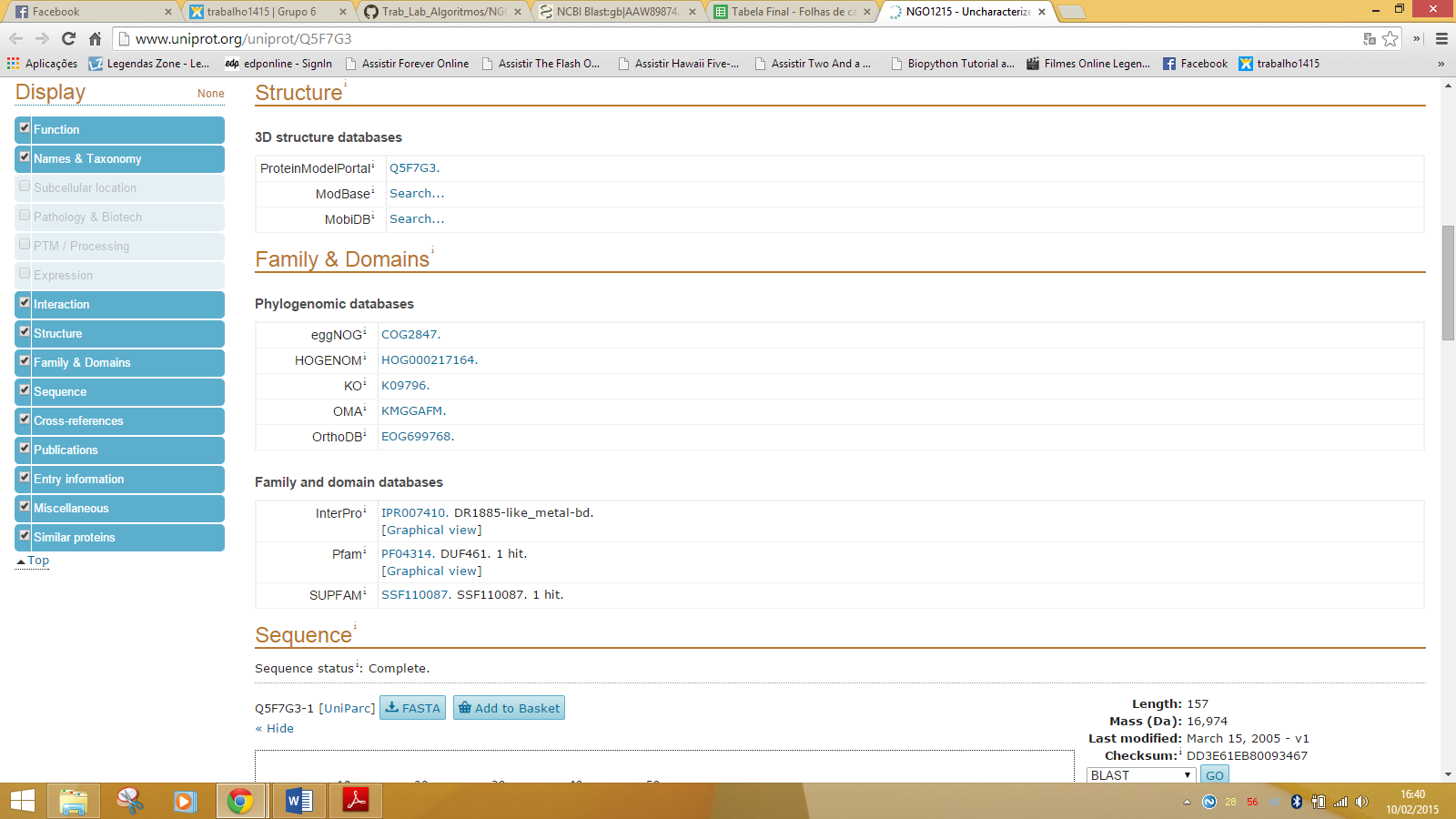






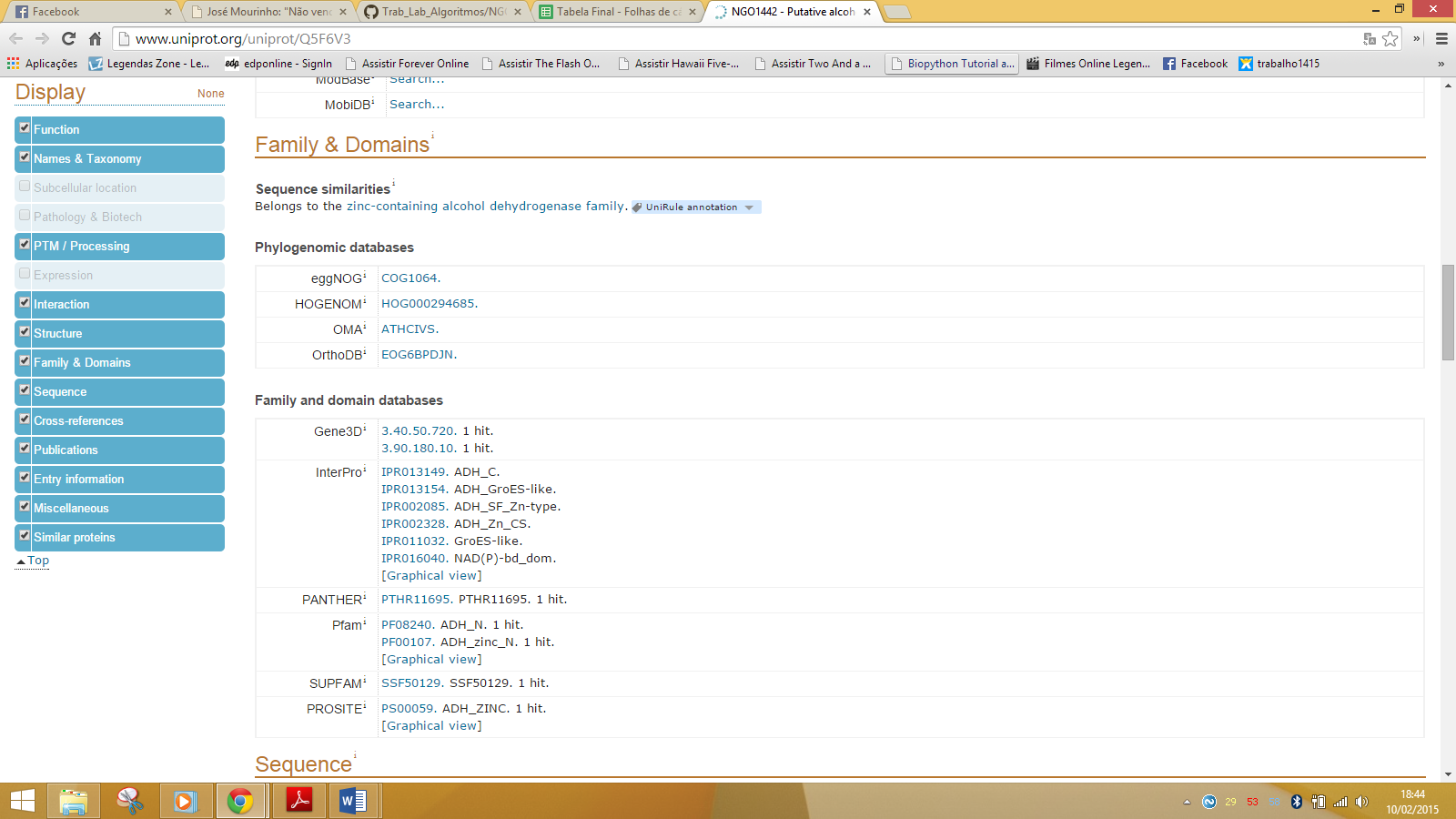
Parte II:

Através do id da proteína vamos para a base de dados Uniprot, e através da família, como dissemos anteriormente, vamos poder inferir a função da proteína.



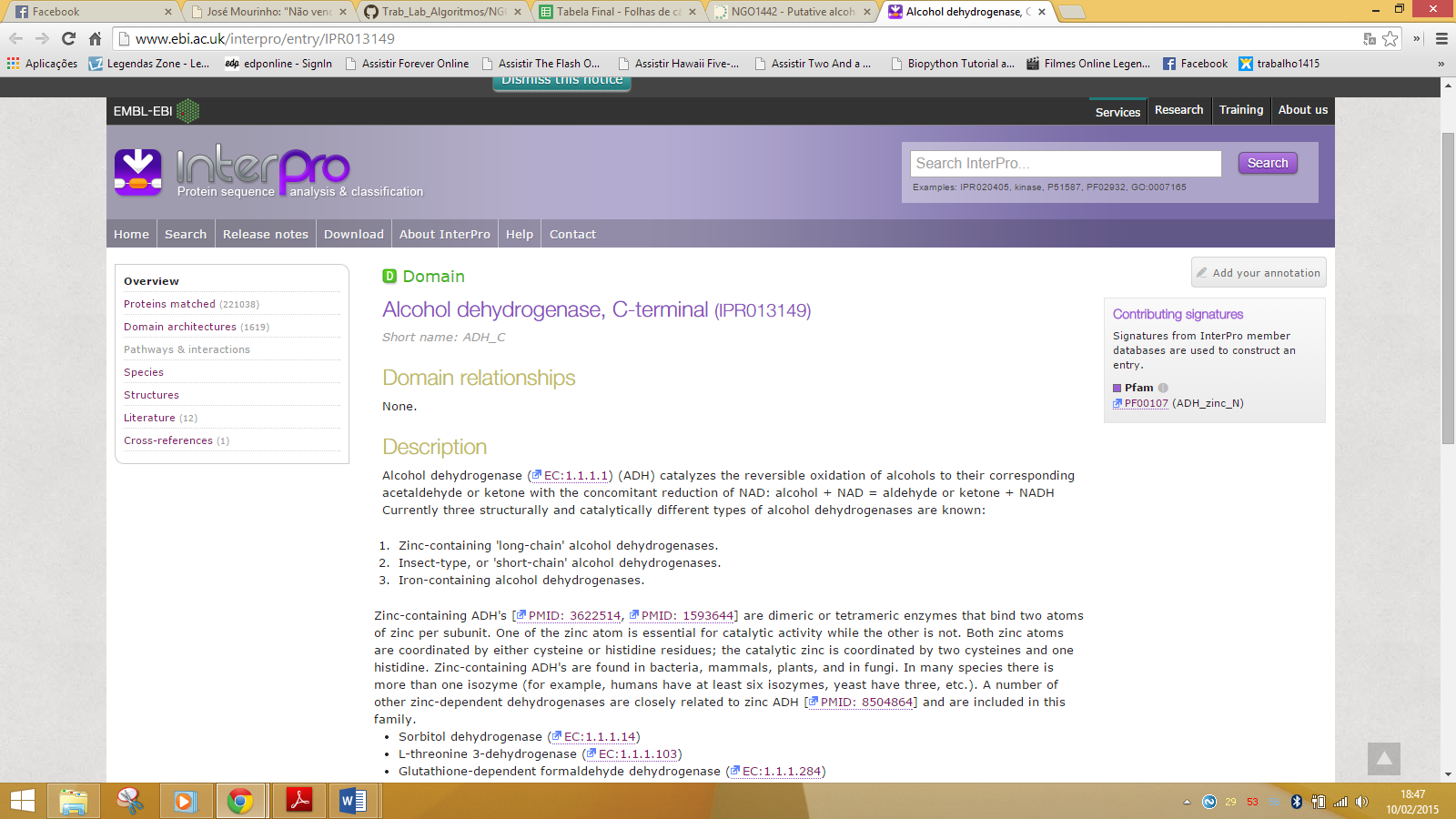
A partir de Family and domain databases, em que temos diversas bases de dados, vamos encontrar diversas informações sendo uma delas a função relativamente a família da nossa proteína.

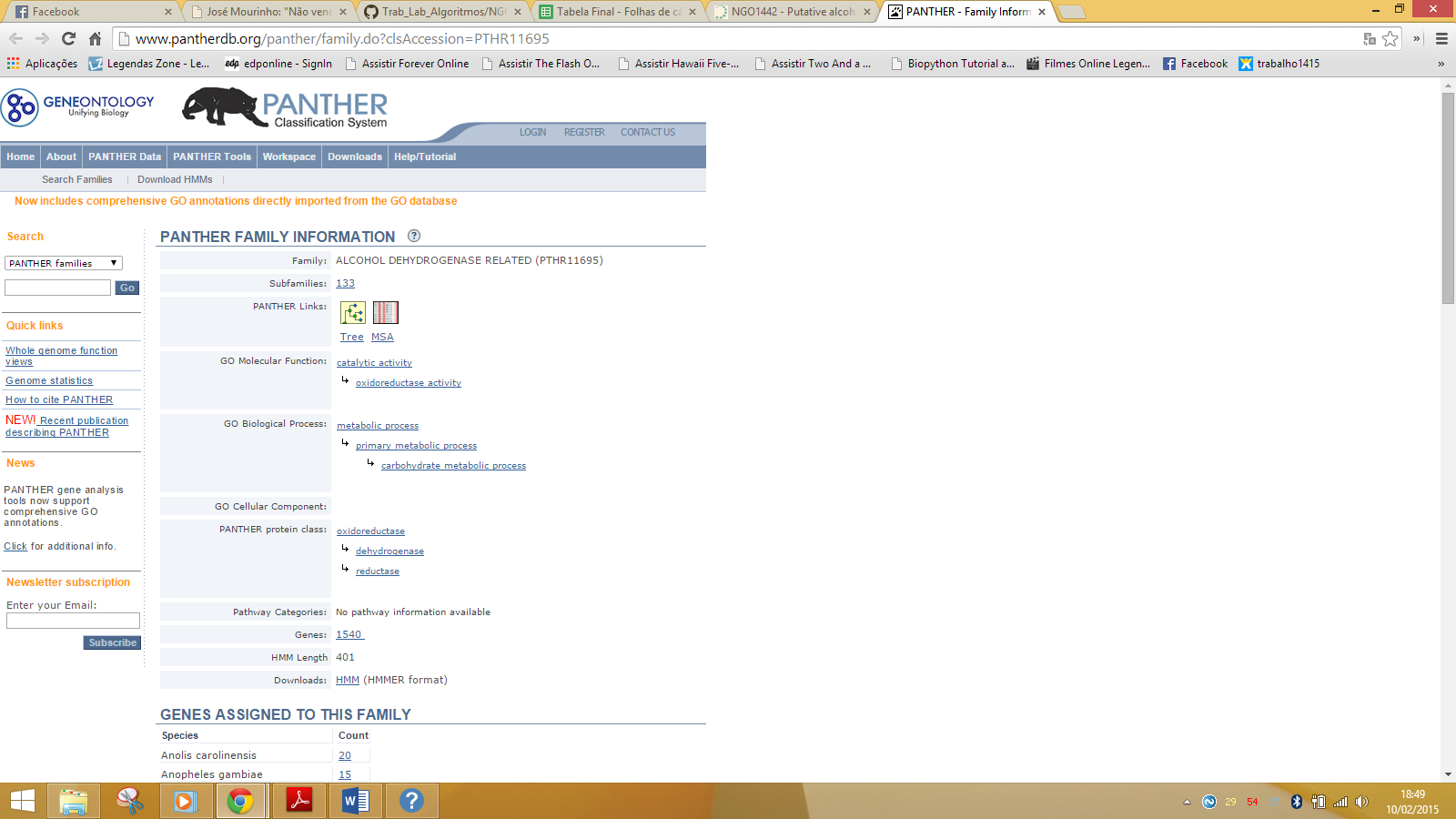
Temos um conjunto de bases de dados que podemos aceder, que dependo da própria podem ser um número maior ou menos, caso esta tenha sido mais estuda ou não.

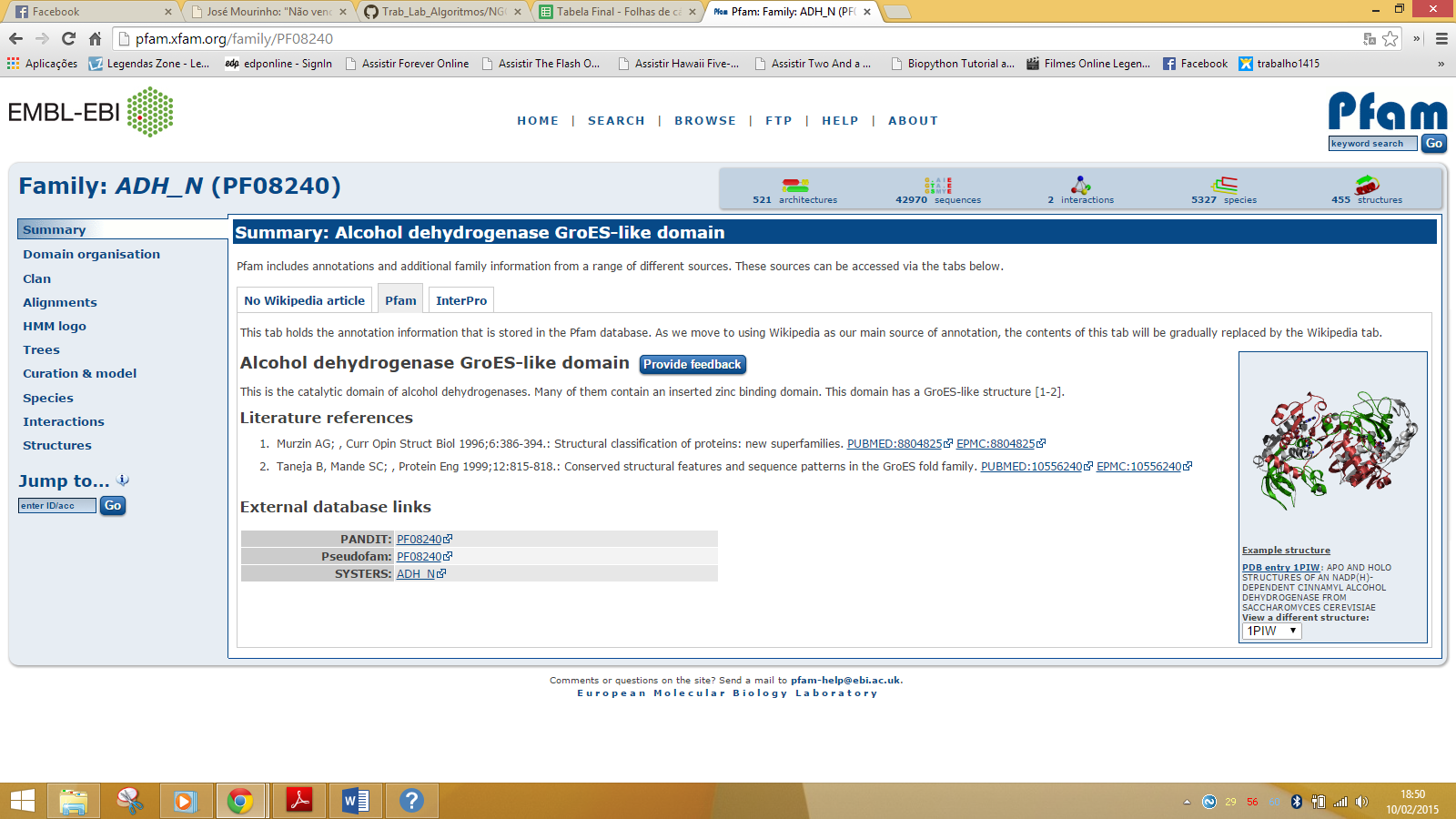


Na imagem de cima, podemos ver um elevado número de base de dados (InterPro, PANTHER, Pfam, SUPFAM, PROSITE). Aqui podemos encontrar diversas informações sobre a função da proteína.

A seguir colocaremos algumas imagens destas referidas proteínas para que seja possível ver como fomos obter a respetiva função (Neste exemplo estamos a usar a proteína com locus\_tag NGO1422).

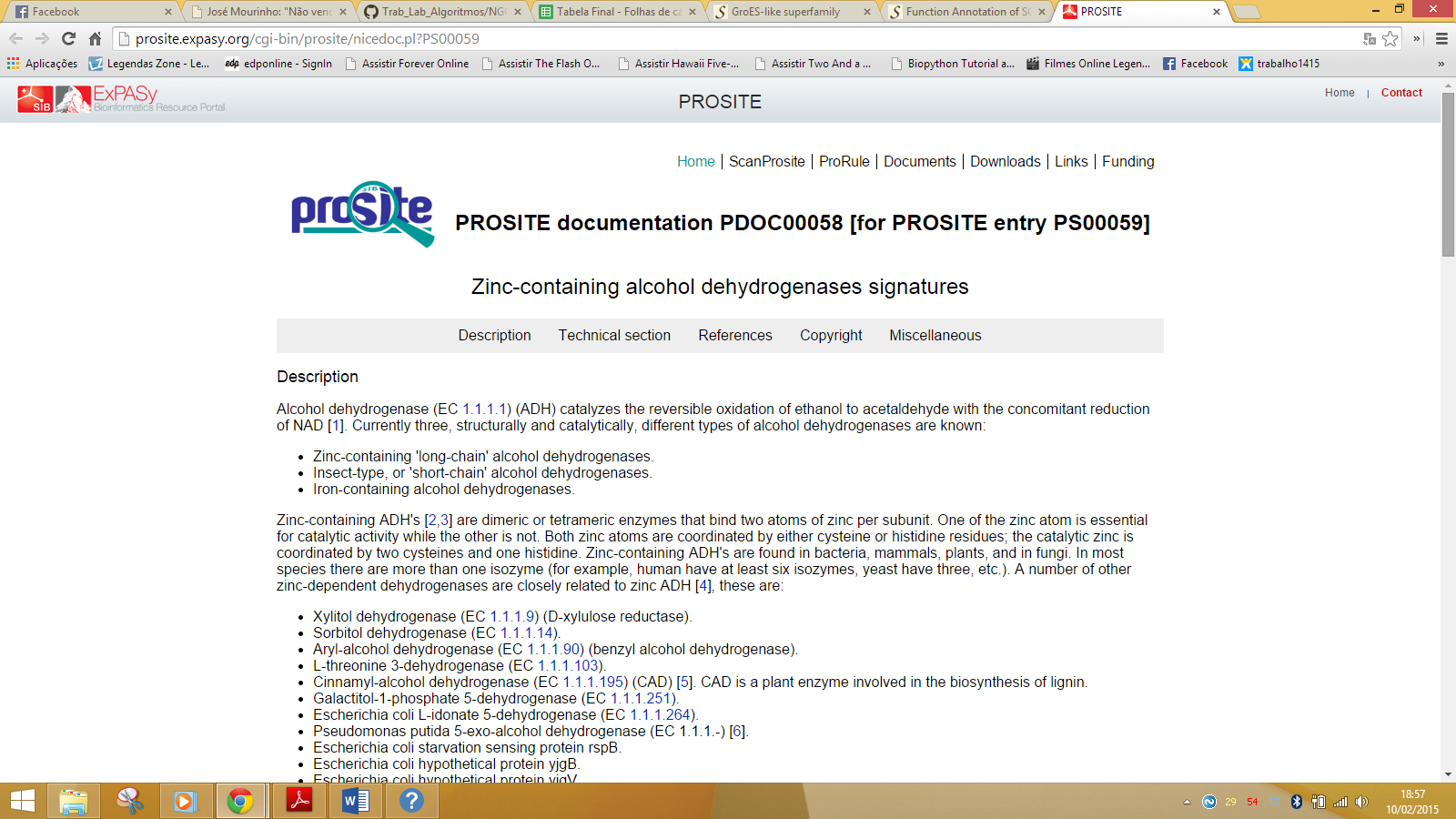
InterPro:

PANTHER:

Pfam:

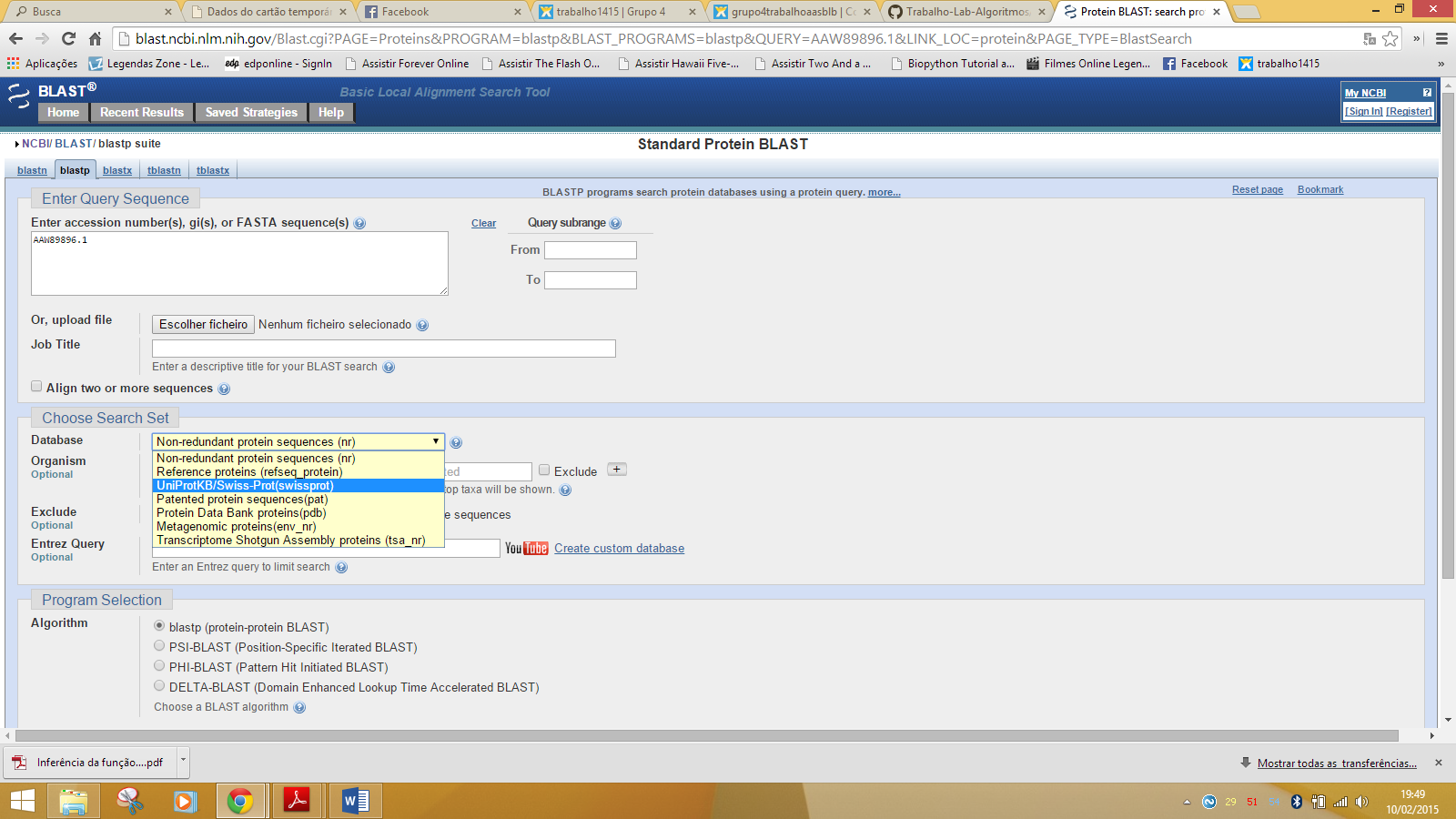
SUPFAM:



PROSITE:

Parte III:

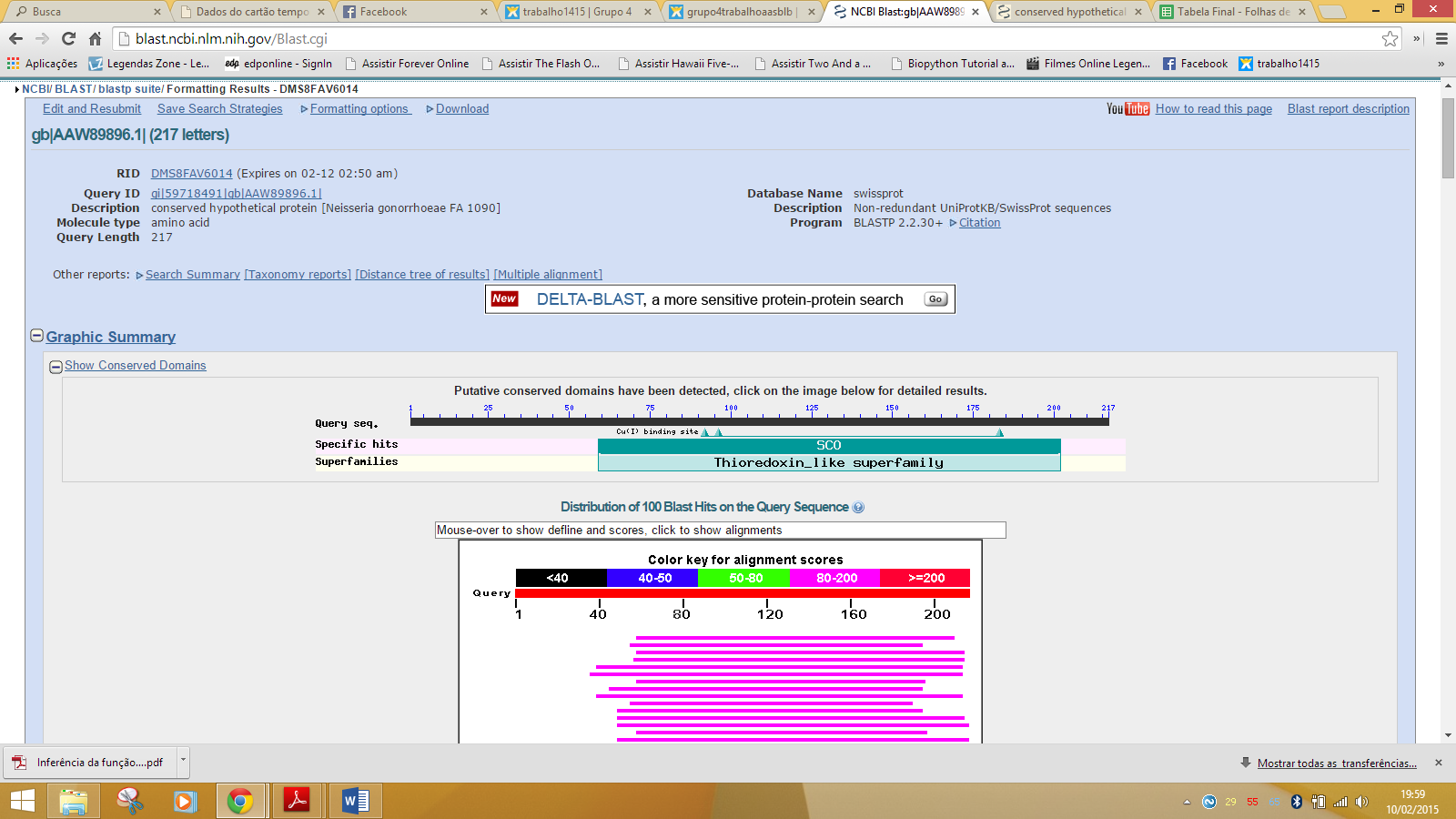
Para a inferirmos a função de proteínas sem função conhecida, recorremos ao BLAST. Através das proteínas homólogas encontradas é possível inferir a função das proteínas em estudo.

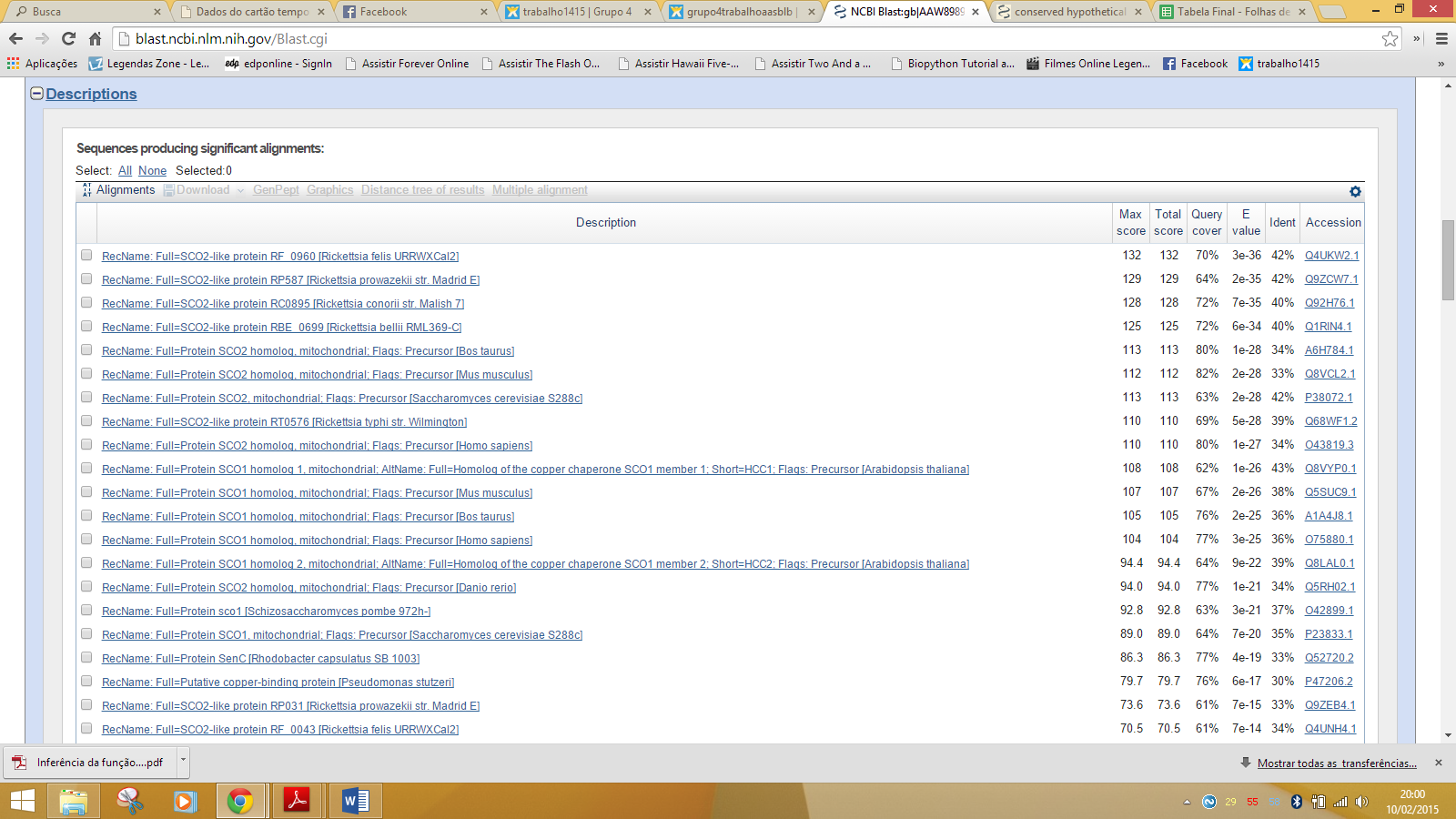
Corremos o BLAST apenas para as proteínas presentes na base de dados SwissProt, fazendo com que os resultados sejam o mais assertivos possível.

Visto termos usado esta base de dados, os resultados que vamos encontrar será em número mais reduzido do que caso tivéssemos efetuado a pesquisa normal (não redundante). Foi usado um limite de e-value de 0,05, ou seja, os resultados apresentados são apenas os que estão abaixo deste valor.

Nos resultados obtidos, nem sempre o valor com menor e-value deve ser o considerado, por esta razão temos que fazer uma análise individual dos resultados para cada proteína, permitindo justificar a escolha do hit em que nos baseamos para a inferência.

Sendo assim, a escolha do melhor resultado será sempre no hit com menor e-value e que ao mesmo tenho tenha uma query mais alta, sendo que podem surgir hits com score alto e bom alinhamento em termos de e-value, mas que no entanto não cobrem a query de forma significativa, comparando com o segundo ou terceiro hit do BLAST.

NGO1237



Bibliografia:

[1]- Watchirs Smith, L.A., et al., Point-of-care tests for the diagnosis of Neisseria gonorrhoeae infection: a systematic review of operational and performance characteristics. Sex Transm Infect, 2013. 89(4): p. 320-6.

[2] Gerbase, A.C., J.T. Rowley, and T.E. Mertens, Global epidemiology of sexually transmitted diseases. Lancet, 1998. 351 Suppl 3: p. 2-4.

[3] Snyder, L.A., S.A. Butcher, and N.J. Saunders, Comparative whole-genome analyses reveal over 100 putative phase-variable genes in the pathogenic Neisseria spp. Microbiology, 2001. 147(Pt 8): p. 2321-32.

[4] CDC. Centers for Disease Control and Prevention 2007. Trends in reportable sexually transmitted diseases in the United States, 2006. http://www.cdc.gov/std/stats/trends2006.htm acessado em

04/02/2015.

[5] Cecil JA, Howell MR, Tawes JJ, Gaydos JC, Mckee Jr KT, Quinn TC, Gaydos CA. Features of

Chlamydia trachomatis and Neisseria gonorrhoeae in Male Army Recruits. J Infect Dis .

2001;184: 1216-19.