

UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES CIÊNCIAS E HUMANIDADES

STÉFFANI VIBANCO DE OLIVEIRA NEVES

**Descoberta de novos epítomos de Trypanosoma cruzi na sua
interação com humanos.**

São Paulo
2022

STÉFFANI VIBANCO DE OLIVEIRA NEVES

**Descoberta de novos epítomos de Trypanosoma cruzi na sua
interação com humanos.**

Versão Original

Trabalho de conclusão de curso apresentado ao curso de Graduação em Biotecnologia da Escola de Artes Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Bacharel em Biotecnologia.

Orientador: Prof. Dr. João Carlos Setubal

Coorientador: Prof. Dr. Luciano Antonio Digiampietri

São Paulo

2022

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico para fins de estudo e pesquisa, desde que citada a fonte.

Dedico esse trabalho a todos professores que tive, que, sem sombra de dúvida, tiveram uma contribuição infinitamente maior para esse resultado, que eu mesma.

AGRADECIMENTOS

Antes de tudo, gostaria de agradecer imensamente ao meu orientador, Prof. Dr. João Carlos Setubal, bem como meu co orientador, Prof. Dr. Luciano Antonio Digiampietri, e principalmente pelo meu supervisor doutorando Gianluca Machado Major da Silva pela paciência e apoio oferecidos a mim durante todo o andamento do projeto, sem ele eu não teria conseguido.

Em seguida, gostaria de agradecer aos mestres, que com tanto carinho nos guiaram durante a graduação, em especial ao Prof. Dr. Tiago Franco, por ensinar que a graduação tem muito mais a oferecer que as aulas teóricas e provas, e que é preciso ter capacidade de equilibrar as responsabilidades com a descontração, ao doutorando Celso Barbiéri, por acompanhar meu trajeto pela graduação sempre com um carinho marsupial. Além disso, quero demonstrar minha gratidão ao Prof. Dr. Felipe Chambergo, e a toda comissão de professores que me acompanharam e me deram todo o suporte necessário para conseguir exercer o meu melhor dentro do curso de Biotecnologia da Universidade de São Paulo.

Por fim, gostaria de agradecer aos verdadeiros amigos, que trilharam esse mesmo árduo caminho juntamente a mim, sempre oferecendo suporte e um ombro amigo nos momentos difíceis, e isso, para mim, teve um papel primordial para minha formação.

RESUMO

Vibanco de Oliveira Neves; Stéffani. Descoberta de novos epítomos de *Trypanosoma cruzi* na sua interação com humanos. Trabalho de conclusão do curso de bacharelado em biotecnologia – Escola de Artes, Ciências e Humanidade, Universidade de São Paulo, São Paulo, 2022.

É de fundamental importância para o desenvolvimento de diversas imunoterapias o mapeamento em larga escala de antígenos e epítomos, mas se torna um grande desafio, especialmente para patógenos eucarióticos, devido aos seus grandes genomas. Neste trabalho, foi desenvolvido um fluxo de processos, a partir de fagos genômicos, para mostrar que bibliotecas imparciais do parasita eucariótico *Trypanosoma cruzi* permitem a identificação de antígenos por amostras de soro de pacientes com doença de Chagas. Foi construída e validada uma biblioteca abrangente da resposta de anticorpos da doença de Chagas, com objetivo de mostrar como os epítomos de conformação linear e putativa (contendo muitos elementos repetidos), permitem que o parasita evite um acúmulo de anticorpos neutralizantes direcionados contra domínios de proteínas que medeiam a patogênese da infecção. Assim, esse fluxo de processos é uma ferramenta reprodutível e eficaz para a identificação de epítomos e antígenos, não apenas para a doença de Chagas, mas talvez também para patógenos emergentes/reemergentes globalmente.

Palavras-chave: Antígenos. Bioinformática. Doença de Chagas. Epítomo. Fago. Imunoterapia.

ABSTRACT

Vibanco de Oliveira Neves; Stefani. Discovery of new epitopes of *Trypanosoma cruzi* in its interaction with humans. Completion work of the bachelor's degree in biotechnology – School of Arts, Sciences and Humanity, University of São Paulo, São Paulo, 2022.

Large-scale mapping of antigens and epitopes is of fundamental importance for the development of several immunotherapies, but it becomes a great challenge, especially for eukaryotic pathogens, due to their large genomes. In this work, a process flow was developed, starting from genomic phages, to show that unbiased libraries of the eukaryotic parasite *Trypanosoma cruzi* allow the identification of antigens by serum samples from patients with Chagas disease. A comprehensive library of the Chagas disease antibody response was constructed and validated, with the aim of showing how epitopes of linear and putative conformation (containing many repeated elements) allow the parasite to avoid an accumulation of neutralizing antibodies directed against protein domains. that mediate the pathogenesis of the infection. Thus, this process flow is a reproducible and effective tool for the identification of epitopes and antigens, not only for Chagas disease, but perhaps also for emerging/re-emerging pathogens globally.

Keywords: Antigens. Bioinformatics. Chagas disease. epitope. Phage. Immunotherapy.

SUMÁRIO

1. INTRODUÇÃO	8
1.1. DOENÇA DE CHAGAS	8
2. OBJETIVOS	12
2.1. OBJETIVO GERAL	12
2.2. OBJETIVOS ESPECÍFICOS	12
3. METODOLOGIA	13
3.1. REVISÃO E ENTENDIMENTO DA LITERATURA	13
3.2. TRATAMENTO DOS DADOS	15
3.2.1. EXTRAINDO INSERTOS A PARTIR DOS READS	15
3.2.2. COMPARAÇÃO DAS SEQUÊNCIAS DE DNA ENTRE SI PARA CONTAGEM DE FREQUÊNCIA	15
3.2.3. REMOÇÃO DE INSERTOS QUE NÃO MAPEIAM NO GENOMA DE T. CRUZI	16
3.3. TRATAMENTO DO PROTEOMA	16
3.3.1. DETERMINAÇÃO DE ORFS DE CADA INSERTO	16
3.3.2. REMOÇÃO DE INSERTOS QUE INICIAM COM FRAME DIFERENTE DE DOIS	16
3.3.3. REMOÇÃO DE INSERTOS QUE NÃO MAPEIAM NO PROTEOMA DE T. CRUZI	17
3.4. VALIDAÇÃO	17
3.5. CLUSTERIZAÇÃO	18
3.6. SEQUÊNCIA DO POSSÍVEL EPÍTOPO, SEQUÊNCIA CONSENSO E VISUALIZAÇÃO NA PROTEÍNA	19
4. RESULTADOS e DISCUSSÃO	19
4.1. HABILIDADES COM LINGUAGENS DE PROGRAMAÇÃO E SISTEMAS DE COMPUTAÇÃO DE ALTO DESEMPENHO	19
4.2. TRATAMENTO DO GENOMA	20
4.3. TRATAMENTO DO PROTEOMA	20
4.4. VALIDAÇÃO	21
4.7. SEQUÊNCIA CONSENSO E VISUALIZAÇÃO NA PROTEÍNA	22
5. CONCLUSÃO	26
6. LIMITAÇÕES DA FERRAMENTA	27
7. REFERÊNCIAS	27
APENDICE A	32
APENDICE B	33

1. INTRODUÇÃO

1.1. DOENÇA DE CHAGAS

A doença de Chagas, também conhecida como tripanossomíase americana, é um distúrbio multissistêmico que pode afetar os sistemas cardiovascular, digestivo e nervoso central.¹ A doença de Chagas é causada pelo *Trypanosoma cruzi*, um parasita hemoflagelado que é transmitido através de várias espécies de insetos reduvídeos hematófagos (barbeiros) principalmente em áreas endêmicas.² A doença foi descrita pela primeira vez por Carlos Chagas em 1909, porém Charles Darwin descreveu seu encontro com o vetor e seus próprios sintomas compatíveis com a doença, indicando que a infecção pelo *T. cruzi* ocorria antes do Dr. Chagas a descrever.³ A Organização Mundial da Saúde (OMS) considera a doença de Chagas uma das vinte doenças tropicais negligenciadas ⁴, e estima-se que 6 a 7 milhões de pessoas estejam infectadas pelo *T. cruzi* em todo o mundo, sendo a grande maioria na América Latina.⁵

A doença de Chagas pode ser considerada uma infecção reemergente, pois áreas em que não havia infecção adquirida localmente estão relatando casos autóctones.⁶ Nos Estados Unidos da América, quando os testes para doença de Chagas em doadores de sangue começaram em 2008, a soropositividade para *T. cruzi* foi de 1 em 6.500 doadores, com 36% deles apresentando evidência clínica de cardiomiopatia chagásica. Em pelo menos 5 desses casos a infecção é evidenciada como tendo ocorrido como transmissão autóctone no Texas.⁷ A doença de Chagas é um importante problema de saúde pública, atinge múltiplos sistemas, incluindo o sistema nervoso central (SNC), o aparelho digestivo, o sistema imunológico e, principalmente, o coração. Na América Latina, ela está entre as causas mais frequentes de insuficiência cardíaca (IC), e é supostamente responsável por até 41% dos casos em áreas endêmicas.⁸

Os estágios do parasita em hospedeiros mamíferos incluem os tripomastigotas da corrente sanguínea e os amastigotos, sem flagelos, replicativos intracelulares, enquanto os estágios no vetor incluem os epimastigotas replicativos e os tripomastigotas metacíclicos infecciosos. As infecções em mamíferos ocorrem quando o parasita está no estágio tripomastigota. Os tripomastigotas infectam uma variedade de células e se convertem em um amastigota replicativo que se multiplica no citoplasma da célula hospedeira. As células parasitadas eventualmente se rompem e liberam tripomastigotas que circulam e podem infectar outras células hospedeiras.⁹

O *T. cruzi* é transmitido em áreas endêmicas por várias espécies de três gêneros de insetos sugadores de sangue triatomíneos, também conhecidos como barbeiros (*Triatoma*,

Panstrongylus, *Rhodnius*).¹⁰ Os três gêneros são amplamente distribuídos na América Latina, do México à Argentina e Chile, e habitam tanto florestas quanto áreas mais secas.¹¹ Entretanto, outros mecanismos de infecção que são importantes, especialmente em áreas não endêmicas, incluem transfusão de sangue, transplante de órgãos, ingestão oral, acidentes laboratoriais, vertical de mãe para filho ou agulhas intravenosas compartilhadas. A transmissão sexual foi relatada por experimentos in vivo com camundongos, mas nenhum relato em humanos está atualmente disponível.¹²

1.2. DIVERSIDADE GENÉTICA

O *T. cruzi* é uma espécie heterogênea com sete cepas, ou unidades de tipagem discretas (DTU), chamadas de TcI, TcII, TcIII, TcIV, TcV, TcVI e Tcbat.¹³ Essa diversidade genética tem sido relacionada à distribuição, patogênese, características clínicas e resposta à terapia. Os parasitas em cada DTU são geneticamente semelhantes e possuem características próximas, incluindo a patologia que causam, bioquímica e imunogenicidade e resistência ao tratamento.¹⁴

O TcI tem uma ampla distribuição, desde o sul dos Estados Unidos até o norte da Argentina e Chile, e esta DTU é a mais frequentemente amostrada em ciclos silvestres, mas também é frequente em ciclos domésticos e é a DTU dominante responsável pela transmissão da doença de Chagas em países endêmicos localizados ao norte da bacia amazônica. Quanto a TcII, V e VI são mais propensos a estar associados a ciclos domésticos e pacientes com doença de Chagas crônica nos países do Cone Sul e Bolívia. As TcIII e IV são principalmente amostradas na floresta tropical. E, por fim, Tcbat previamente identificado em morcegos, foi recentemente encontrado em humanos. É bem conhecido que várias DTUs podem coexistir no mesmo vetor e em um único hospedeiro.¹⁵

A variedade genética apresentada por esse parasita também fica evidente quando se analisa as famílias de multigenes que ele possui. Isso ocorre porque o genoma do *T. cruzi* possui muitas sequências repetidas, indicando que muitos desses genes estão em desequilíbrio de ligação e que ocorre reprodução clonal dessa população. Ainda, a presença de famílias de multigenes está relacionada com a capacidade dele de invadir células e apresentar tropismo para diferentes tecidos, causando diferentes tipos de cardiopatias e mega síndromes associadas à doença de Chagas. Desse modo, essa variedade genética está associada com a infectividade do *T. cruzi*, pois muitas dessas famílias codificam para genes presentes na superfície do protozoário, como trans-sialidases e mucinas.

1.3. AUTOIMUNIDADE NA DOENÇA DE CHAGAS

O *T. cruzi* apresenta diferentes estratégias de fuga que permitem sua evasão do sistema imune do hospedeiro, possibilitando sua persistência e o estabelecimento de infecção crônica que leva ao desenvolvimento da cardiomiopatia chagásica crônica (CCC). Os potentes estímulos imunes gerados pela persistência do *T. cruzi* podem resultar em dano tecidual e resposta inflamatória. Além disso, o mimetismo molecular entre as moléculas do parasita e as proteínas do hospedeiro pode resultar em reação cruzada com moléculas próprias e, conseqüentemente, em características autoimunes, incluindo auto-anticorpos e células autorreativas. Embora controversa, há evidências que demonstram um papel da autoimunidade na progressão clínica da CCC. No entanto, o mecanismo exato subjacente à geração de uma resposta autoimune na progressão da doença de Chagas humana é desconhecido.¹⁶

Há dois mecanismos que tentam explicar a autoimunidade em Chagas: uma com a ativação de linfócitos B e T de forma antígeno independente, e a outra o mimetismo molecular. O primeiro mecanismo, junto com a presença dos antígenos do parasita, pode desencadear em um grande dano tecidual, ultrapassando o limiar da autotolerância e induzindo a produção de autoanticorpos. O mimetismo, por sua vez, se dá pela semelhança entre antígenos do *T. cruzi* e humanos, e assim ocorre reação cruzada de anticorpos.¹⁶

1.4. PHAGE DISPLAY

A técnica de *phage display* ou exibição de fagos foi inicialmente desenvolvida e usada para mapear epítomos de anticorpos¹⁷, identificar antígenos envolvidos em doenças como câncer¹⁸, doenças¹⁹, e infecções parasitárias²⁰, incluindo a doença de Chagas.²¹ O Phage Display envolve inserir um fragmento de DNA em um bacteriófago geneticamente modificado, que expressa em seu capsídeo viral um peptídeo, de modo que o peptídeo (ou anticorpo) correspondente codificado pelo fragmento de DNA exógeno seja exibido na superfície do bacteriófago. Se o peptídeo é um fragmento de um antígeno reconhecido por um dado anticorpo, a partícula do bacteriófago pode ser capturada, a partir da exposição da proteína que interage com o alvo ligante²². Neste processo de seleção de afinidade chamado biopanning, a biblioteca é então apresentada a moléculas alvo, normalmente imobilizadas em suportes sólidos. As interações fracas entre o fago que expressa a proteína e o alvo são interrompidas por lavagens sucessivas, enquanto os fagos que contêm moléculas com alta

afinidade pelo alvo são recuperados por eluição.²³ E, assim, o antígeno é isolado do conjunto de partículas de fago.²⁴

1.5. IEDB (BANCO DE DADOS DE EPÍTOPOS IMUNES)

O banco de dados de epítomos imunes (IEDB) é um serviço gratuito com o objetivo de auxiliar pesquisas imunológicas. Nele é possível encontrar resultados de mais de 1,6 milhões de experimentos de resposta imune adaptativa a epítomos, reunidos principalmente na literatura.²⁵ Esses dados vêm de 19.500 publicações, incluindo toda a literatura disponível desde o início do PubMed até o presente. As pesquisas são realizadas no PubMed a cada duas semanas permitindo uma atualização com novos conteúdos.²⁵

O IEDB possui uma grande relevância para este projeto, é a partir dele que se torna possível uma validação de que nossos dados se tratam de possíveis epítomos, uma vez que é muito provável que encontremos epítomos já conhecidos. Além disso, é possível ter um melhor entendimento das características de um epítomo por meio de um vasto banco de dados.

1.6. CONTEXTO DESTE TRABALHO

Este trabalho de conclusão de curso faz parte de um projeto colaborativo entre o orientador Professor Doutor João Carlos Setubal e o Professor Doutor Ricardo Giordano. Esse projeto, liderado pelo Prof. Ricardo, tem como objetivo identificar epítomos do *T. cruzi* por meio de phage display. Nesse projeto, até o momento foram analisados 8 conjuntos de dados ([Tabela 1](#)).

Foi publicado uma análise preliminar sobre esses conjuntos de dados no artigo “A refined genome phage display methodology delineates the human antibody response in patients with Chagas disease”²⁴ de Teixeira et al.,

Este trabalho de conclusão de curso tem o objetivo de refinar a metodologia utilizada no artigo de Teixeira et al., para um conjunto de dados específico, o conjunto CCC_mild A e B. Este trabalho também está associado ao trabalho de doutorado do aluno Gianluca Machado da Silva, orientado pelos profs. Setubal e Giordano. Gianluca foi co-supervisor deste trabalho de TCC.

2. OBJETIVOS

2.1. OBJETIVO GERAL

Geração de uma nova lista de potenciais epítomos na interação *T. cruzi*-humanos, a partir da análise de um conjunto de dados obtidos pela técnica de Phage Display.

2.2. OBJETIVOS ESPECÍFICOS

2.2.1. Familiarização com os seguintes temas:

- Doença de Chagas e *Trypanosoma cruzi*;
- Técnica de Phage display;
- Conceito de epítopo;
- Ambiente computacional do setulab;
- Alinhamento de sequências;
- Clusterização de sequências;
- BLAST (*Basic Local Alignment Search Tool*);
- IEDB [25](#) .

2.2.2. Para os conjuntos de dados CCC_mild A e B de phage display:

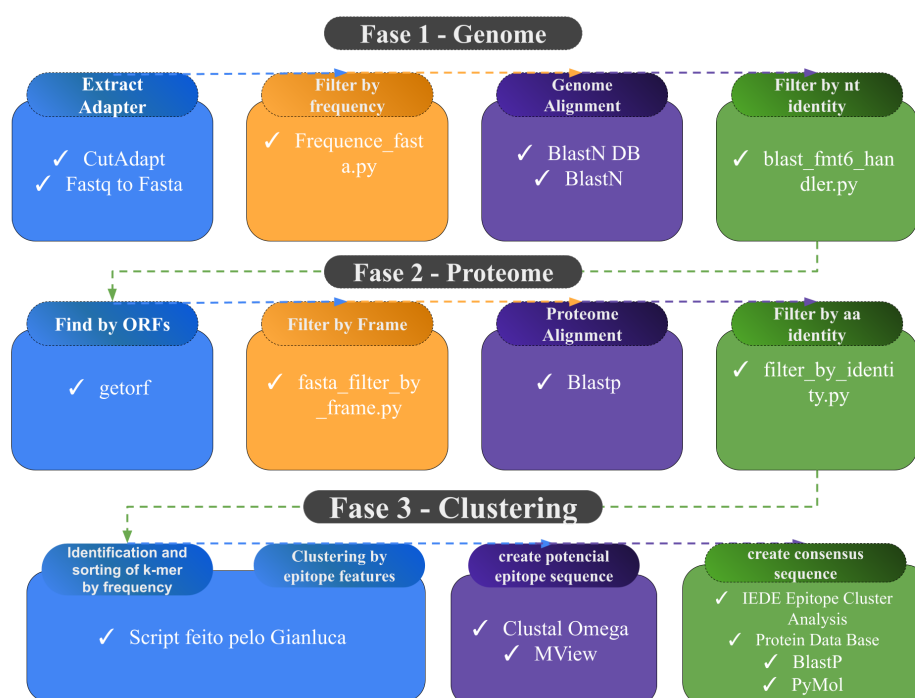
- Extrair insertos a partir dos reads;
- Comparação das sequências de DNA entre si para contagem de frequência
- Remover insertos que não mapeiam no genoma de *T. cruzi*;
- Determinação das ORFs (*Open Read Frame*) de cada inserto;
- Comparação das ORFs entre si para remover duplicações;
- Comparação das ORFs resultantes com proteoma de *T. cruzi*;

- Comparação das ORFs resultantes com epítomos de *T. cruzi* em IEDB.

3. METODOLOGIA

O estudo foi dividido em 5 etapas: Revisão e Entendimento da literatura (3.1); Tratamento do Genoma (3.2); Tratamento do Proteoma(3.3); Validação (3.4), Clusterização(3.5) e Sequência Consenso e Visualização na proteína (3.6). Dentre essas etapas, às 3.2, 3.3, 3.5 e 3.6 formam um pipeline para a identificação de epítomos a partir de sequências de antígenos, conforme figura 1.

Figura 1: Diagrama de fluxo de etapas do pipeline para identificação de epítomos a partir de sequências de antígenos obtidas por meio de técnica de Phage Display.



Fonte: Stéffani Vibanco de Oliveira Neves (2022).

3.1. REVISÃO E ENTENDIMENTO DA LITERATURA

Para obter-se um melhor entendimento dos processos e temáticas que foram utilizados foi realizado um levantamento de literatura científica.

O entendimento se baseou fundamentalmente no artigo “A refined genome phage display methodology delineates the human antibody response in patients with Chagas disease” de Teixeira et al.²⁴, e também no artigo “Protocol for design, construction, and selection of genome phage (gPhage) display libraries.” de Rodriguez Carnero et al.²⁶

Decidimos abordar a problemática a partir de um conjunto de dados provenientes de uma biblioteca de Phage Display (Tabela 1) advinda do artigo de referência.²⁴ Com intenção de obter-se uma maior variedade genética, foram utilizadas pelo autor, como fonte para essa biblioteca, amostras de soro de pacientes contaminados pela doença de Chagas com diferentes níveis de sintomas, e que se enquadram nas seguintes exigências.²⁴

- a. Pacientes com pelo menos dois resultados positivos para a presença de anticorpos anti-T. cruzi.
- b. Todos os pacientes candidatos foram submetidos a eletrocardiograma (ECG) e ecocardiograma (ECO) e aqueles com ECG anormal foram classificados como tendo cardiomiopatia leve, quando a fração de ejeção do ventrículo esquerdo (FEVE) foi superior a 40% ($FEVE > 40\%$), ou cardiomiopatia grave, quando a fração de ejeção do ventrículo esquerdo foi menor ou igual a 40% ($FEVE \leq 40\%$).
- c. Os pacientes sem alterações eletrocardiográficas foram considerados assintomáticos.
- d. As amostras de soro foram agrupadas em grupos de 10 doadores para formar dois conjuntos independentes (duplicatas biológicas) para cada condição de doença:
 - i. controle (2 x 10 doadores)
 - ii. assintomático (2 x 10 doadores)
 - iii. cardiomiopatia leve (2 x 10 doadores)
 - iv. cardiomiopatia grave (2 x 10 doadores).

Tabela 1: Conjuntos de reads disponíveis. Para cada conjunto (linha) há dois lotes (A e B)

control	pacientes controle
asympto	pacientes com teste positivo para doença de chagas mas sem sintomas
CCC_mild	pacientes com teste positivo para doença de chagas com sintomas leves
CCC_severe	pacientes com teste positivo para doença de chagas com sintomas graves

Fonte: Stéffani Vibanco de Oliveira Neves , 2022²⁴

Neste projeto foi utilizado o conjunto CCC_mild_A (também chamado de K) e B (também chamado de conjunto O).

3.2. TRATAMENTO DOS DADOS

3.2.1. EXTRAINDO INSERTOS A PARTIR DOS READS

A partir do descarregamento do conjunto de dados obtidos pelo sequenciamento, se fez necessário a retirada das sequências de adaptadores, esses adaptadores contêm os índices (sequência curta de bases que identificam cada amostra) e ficam no início e no fim da sequência. Foi identificado que para esse sequenciamento, a sequência de adaptadores a ser retirada é de upstream “ATGACCATGGCAGTAC” e downstream “GTACCCGGTGCGCCGG” e para a retirada, foi utilizada a ferramenta de linha de comando Cutadapt.²⁷

Além disso, foi utilizado o conversor, de linha de comando, de arquivo de fastq para fasta. Dessa forma obtemos os dados contendo apenas a sequência de interesse inicial contendo os possíveis antígenos de *T. cruzi*.

3.2.2. COMPARAÇÃO DAS SEQUÊNCIAS DE DNA ENTRE SI PARA CONTAGEM DE FREQUÊNCIA

As sequências que fazem parte do conjunto possuem grande diversidade. Para garantir uma maior confiabilidade foi definido que elas precisam ter uma frequência mínima de duas aparições, essa informação foi observada no artigo de Teixeira et al., que cita a referência de Dias-Neto et al, 2009.²⁸ Para esse filtro, foi utilizado o script em Python²⁹ que verificou a aparição das sequências repetidas e criou um novo arquivo contendo apenas as sequências com uma frequência mínima de dois.

3.2.3. REMOÇÃO DE INSERTOS QUE NÃO MAPEIAM NO GENOMA DE *T. CRUZI*

A fim de identificar as sequências que representam o genoma do *T. cruzi*, foi realizado um alinhamento das possíveis sequências de antígenos com as sequências de *T. cruzi* (cepas CL Brener, Sylvio X10, DM28c, and Marinkellei descarregadas do NCBI³⁰). Para isso, foi utilizado o aplicativo de linha de comando blastN³¹, com os parâmetros: número de alinhamentos (num_alignments) em dez e número máximo de HSPs (High-Scoring Segment Pairs) em 1.

Após o alinhamento foi utilizado um script em Python³² para verificar a porcentagem de identidade de cada sequência com as sequências de *T. cruzi* e descartar as que obtiveram um valor menor de 90%. O ideal seria que utilizássemos um valor de 100% de identidade, porém devido às variações de cepas do *T. cruzi* das amostras, e a utilizada para o alinhamento, houve uma maior tolerância a sua identidade.

3.3. TRATAMENTO DO PROTEOMA

3.3.1. DETERMINAÇÃO DE ORFS DE CADA INSERTO

As sequências que foram obtidas até agora são constituídas por nucleotídeos, então para a construção de um proteoma foi necessária a utilização de um programa de linha de comando chamado EMBOSS GetOrf³³, com ele, foi possível obter sequências de quadros de leitura aberta (ORFs).

3.3.2. REMOÇÃO DE INSERTOS QUE INICIAM COM FRAME DIFERENTE DE DOIS

Os insertos proveniente do fago de DNA supõe que o quadro correto para a tradução de aminoácidos é o quadro dois positivo (+2). Portanto, nessa etapa se fez necessário descartar os quadros de leitura aberta (ORF) que não iniciasse no quadro +2, para isso foi utilizado um script em Python³⁴ que filtrou as sequências de ORFs que estavam no quadro +2, criando um arquivo apenas com as desejadas.

3.3.3. REMOÇÃO DE INSERTOS QUE NÃO MAPEIAM NO PROTEOMA DE *T. CRUZI*

A fim de identificar as sequências que representam o proteoma do *T. cruzi*, foi realizado um alinhamento das possíveis sequências de antígenos com as sequências de *T. cruzi* (cepas CL Brener, Sylvio X10, DM28c, and Marinkellei descarregadas do NCBI³⁰). Para isso, foi utilizado o aplicativo de linha de comando blastP³¹, com os parâmetros: número de alinhamentos (num_alignments) em dez e número máximo de alinhamentos (max_hsps) em 1, o tamanho das sequências de correspondência (word_size) de valor 6, número de aberturas (gapopen) em 13 e o número de alinhamentos que seriam esperados (evaluate) em 100.

Após o alinhamento foi utilizado um script em Python³⁵ para verificar a porcentagem de identidade de cada sequência com as sequências de *T. cruzi* e descartar as que obtiveram um valor menor de 70%. Este valor se deve ao filtro de 90% feito na etapa 3.2.3, uma vez que ele foi utilizado para o filtro de identidade de nucleotídeos, surge uma necessidade de uma maior tolerância quando tratamos de peptídeos. Essa afirmação se justifica pela ocorrência de famílias gênicas, que podem ter proteínas semelhantes codificadas por genes diferentes. Sabe-se que no genoma do *T. cruzi* existem famílias gênicas, e portanto, usando essa tolerância, foi possível fazer um mapeamento de peptídeos traduzidos no proteoma levando em conta esse fenômeno.

3.4. VALIDAÇÃO

Com a intenção de validar que as sequências que passaram pelo pipeline contém epítomos de *Trypanossoma cruzi*, nesta fase do pipeline foi descarregado todas as sequências de epítomos que estão presentes no banco de dados do website IEDB^{referencia}, com os seguintes parâmetros: Epitope Linear, Organismo *Trypanossoma cruzi*, Hospedeiro Humano, doença de Chagas e outros parâmetros com valores padrões já inseridos pela plataforma.

Após o descarregamento, foi criado um banco de dados Blast com essas sequências, e assim foram alinhadas as sequências de epítomos do IEDB com as possíveis sequências de antígenos dos fluxos anteriores, para isso foi utilizado o aplicativo de linha de comando BlastP.³¹

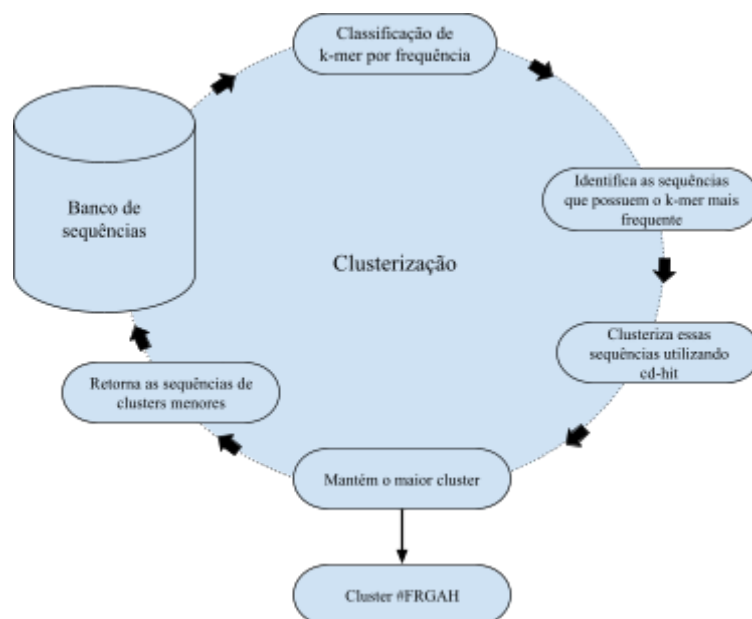
3.5. CLUSTERIZAÇÃO

A fim de obter-se uma análise otimizada, foi decidido agrupar as sequências resultantes da etapa “3.3 tratamento do proteoma” em clusters. Para isso, foi estruturado um script em Python³⁶ que, a partir do banco de sequências, detecta subsequências (k-mer) de tamanho oito de cada sequência, e ordena esses k-mers, em ordem decrescente, conforme sua frequência de aparição, da maior para a menor. Então, ele usa o aplicativo cd-hit³⁷, que tem como objetivo clusterizar sequências de proteínas com ao menos 80% de identidade, utilizando as sequências que possuem a aparição deste k-mer mais frequente.

A partir disso, obtemos diferentes clusters advindos das sequências utilizadas. O cluster que obteve um maior número de sequências agrupadas será considerado para análise final, enquanto as sequências que fazem parte de clusters de menor número voltarão ao banco de sequências para serem reconsideradas por outro k-mer, seguindo a ordem de frequência deles, conforme [figura 2](#).

Como resultado desse script, foram criadas pastas para cada sequência de k-mer em que dentro delas, há os clusters criados pelo cd-hit, incluindo o cluster com um maior número de sequências. Além disso, foi criado um arquivo de log que traz informações sobre as ocorrências do script, um arquivo contendo informações sobre os maiores clusters de cada k-mer.

Figura 2: Processo de clusterização a partir de script Python



Fonte: Stéffani Vibanco de Oliveira Neves (2022).

3.6. SEQUÊNCIA DO POSSÍVEL EPÍTOPO, SEQUÊNCIA CONSENSO E VISUALIZAÇÃO NA PROTEÍNA

A partir do documento que contém informações sobre os maiores clusters de cada k-mer, foi selecionado um cluster em que seu possível epítipo já estivesse catalogado no banco de dados IEDB.

Então, a fim de obter-se uma sequência que possivelmente representa um epítipo deste cluster, foi utilizado o programa Clustal Omega³⁸ contendo as sequências completas dele. O resultado foi visualizado pelo outro aplicativo, o Mview³⁸. Então, com o objetivo de verificar se esse possível epítipo já está catalogado no IEDB, procuramos essa sequência no resultado do BlastP com o IEDB (feito na etapa “3.4 - Validação”).

Por outro lado, para obter-se uma sequência consenso deste cluster, foi utilizado com os parâmetros padrão, o web programa Epitope Cluster Analysis³⁹, contendo o fasta com sequências completas dele. O resultado foi a sequência consenso das sequências.

Ao obter a sequência consenso de 100% deste cluster, foi utilizada a ferramenta BlastP³¹ do NCBI, utilizando diversos bancos de dados de proteínas, para identificar qual proteína do organismo *Trypanossoma cruzi* poderia corresponder a esta

sequência consenso. Então, foi selecionada a proteína que continha uma maior correspondência com o possível epítipo.

A partir disso, essa proteína foi encontrada no UniProt⁴⁰, um banco de dados de proteínas que contém além de proteínas com estruturas já validadas, previsões das não validadas. Então, utilizamos o aplicativo Desktop PyMol⁴¹ para a visualização desta proteína, e, por conseguinte, foi utilizado a função de “seleção de sequência de resíduos de proteína” com código de apenas uma letra para selecionar a parte da proteína em que houve um alinhamento com a sequência consenso, essa parte foi colorida para uma melhor visualização.

4. RESULTADOS e DISCUSSÃO

4.1. HABILIDADES COM LINGUAGENS DE PROGRAMAÇÃO E SISTEMAS DE COMPUTAÇÃO DE ALTO DESEMPENHO

Foram realizados scripts em Python que usam os módulos de `argparse`⁴², que possibilitam a criação de interfaces de usuário na linha de comando. Além disso, utilizou-se módulos da biblioteca de `biopython`⁴³, que é uma biblioteca de ferramentas para dados biológicos, módulos da biblioteca de `pandas`⁴⁴ que é utilizada para manipulação e análise de dados. Os scripts são armazenados no github do doutorando Gianluca Major e da própria autora, e estão sendo produzidos ao longo do projeto de doutorado do doutorando Gianluca Major.

Ainda, foi utilizada a ferramenta de desenho do Google Drive⁴⁵ para confeccionar infográficos que facilitaram a compreensão dos fluxos de processos neste projeto.

4.2. TRATAMENTO DO GENOMA

As sequências obtidas a partir da biblioteca foram advindas de pacientes com teste positivo para doença de chagas com sintomas leves. O grupo de sequências CCC_mild_A possui 868 mil sequências, enquanto o grupo de CCC_mild_B possui 651 mil sequências. Conforme as sequências caminham pelo pipeline proposto, ocorre uma alteração no número de sequências em cada grupo, devido a filtros e alinhamentos que proporcionam sequências cada vez mais próximas de epítopos do organismo *Trypanosoma cruzi*, essas alterações podem ser vistas na [tabela 2](#).

Inicialmente, foram retirados os adaptadores de cada sequência, assim foi possível obter apenas os nucleotídeos de interesse que representam os possíveis

antígenos, por conseguinte foram eliminadas as sequências que são únicas, ou seja, apresentaram apenas uma recorrência em todo banco de sequências, e assim foram reduzidos erros do processo de phage display, aumentando a confiança dos dados. Nessa etapa, a quantidade de sequências do grupo A e B foram, respectivamente, 48435 e 21154 sequências.

Então, após alinhamento entre as sequências dos grupos A e B, com as sequências de *Trypanossoma cruzi* já conhecidas, provindas pelo National Center for Biotechnology Information (NCBI) e a filtragem por identidade de 90%, foram obtidas tanto no grupo A, quanto no B sequências ([Tabela 2](#)) que possuem uma maior confiabilidade de pertencerem ao organismo de interesse.

4.3. TRATAMENTO DO PROTEOMA

As sequências, até então de nucleotídeos, foram transformadas em fases de leitura aberta (Open Reading Frames - ORF), obtendo-se sequências entre os códons de início e parada, quanto a quantidade de sequências do grupo A e B foram, respectivamente, 414251 e 175888, conforme [tabela 2](#). Após isso, ocorreu um filtro em que foram descartadas as ORFs que não se iniciavam no quadro +2, o que resultou numa grande diminuição no número de sequências dos grupos A e B, respectivamente. 35231 e 16873.

Então, após o alinhamento entre os proteomas dos grupos A e B, com os proteomas de *Trypanossoma cruzi* já conhecidos provindas do NCBI, e uma filtragem por identidade de 70%, obtivemos 13569 sequências no grupo A, e 12162 sequências no grupo B.

Tabela 2: Sequências de possíveis antígenos de *Trypanosoma cruzi* conforme sua passagem no pipeline.

Etapas	ccc_mild_a (K)	ccc_mild_n (O)
Sequências Originais	868305	651503
Exclusão de Sequências Únicas	48435	21154
Sequências com 90% de Identidade após BlastN	48118	21138
Orfs que iniciam com frame 2	35231	16873
Sequências com 70% de Identidade após BlastP	13569	13569

Fonte: Stéffani Vibanco de Oliveira Neves , 2022.

4.4. VALIDAÇÃO

O banco de dados de sequências dos grupos A e B até agora possuem sequências com possíveis antígenos do *Trypanosoma cruzi*, então foi feito um alinhamento com as sequências de epítomos que estão presentes no IEDB. Conforme a tabela 3, podemos perceber que das sequências do no grupo A que passaram pelo pipeline proposto, há 14,68% possuem ao menos um epítomo já conhecido pelo IEDB enquanto no grupo B, há 77,35%. Ao excluir os epítomos repetidos obtemos a quantidade de epítomos únicos do IEDB que foram encontrados nas sequências deste projeto. ([Tabela 3](#))

Tabela 3: Sequências de possíveis antígenos de *Trypanosoma cruzi* conforme sua passagem no pipeline.

Descrição	Total de Antígenos	Alinhamento com ao menos um epítomo do IEDB com 100% de identidade	Número de epítomos únicos no IEDB
CCC_mild_a	13569	1993	376
CCC_mild_b	12162	9408	354

Fonte: Stéffani Vibanco de Oliveira Neves , 2022.

4.5. CLUSTERIZAÇÃO

Ao passarmos o banco de sequências desse projeto pelo script de clusterização explicado na Seção 3.5, obtivemos 637 clusters para o grupo A, e 66 para o grupo B, cada um com sua sequência representativa. Com esse resultado, podemos observar que no grupo B, há uma maior semelhança entre as sequências que contém, do que o grupo A. Como o número de epítomos do IEDB que foram identificados para os dois grupos é semelhante, afinal são as mesmas amostras biológicas, este resultado sugere que no grupo A haja muitas sequências que não são de antígenos. Com efeito, no artigo de Teixeira et al., se menciona que essa divergência pode ser devido às diferenças no “input” do fago no phage display, que foi menor para a seleção do grupo B. No entanto, 70% de todos os epítomos do grupo B estavam contidos no A, confirmando assim a confiabilidade do processo de seleção usado aqui.

4.6. POSSÍVEIS EPÍTOPOS

Ao final da clusterização, obtivemos uma lista de possíveis epítomos, tanto do grupo A quanto do grupo B. Essa lista se encontra respectivamente nos apêndices A e B do trabalho.

Nesses apêndices há uma tabela com diversos dados que se fazem relevantes para a seleção dos clusters e possíveis epítomos. Na primeira coluna da tabela encontrada nos apêndices, é o valor do **K-mer**, que possui um tamanho escolhido de 8 aminoácidos que foi listado conforme sua frequência, a próxima coluna é essa **frequência**, então a coluna seguinte representa as **sequências únicas** que possuem o k-mer, após isso a coluna posterior representa a quantidade de sequências que estão presentes no **maior cluster** desse grupo de k-mer. A coluna consecutiva possui 4 valores, representados pela quantidade de sequências de **menor frequência**, a de **maior frequência**, o **tamanho médio** e o **desvio padrão**. Além disso, nas colunas seguintes encontram-se, o **número de identificação do maior Cluster** desse grupo de k-mer, o **número de identificação da sequência representativa** do maior cluster, e a própria **sequência representativa**.

4.7. SEQUÊNCIA CONSENSO E VISUALIZAÇÃO NA PROTEÍNA

Ao obter-se os maiores clusters de cada grupo de k-mer, foi escolhido um cluster que seu possível epitopo já estivesse catalogado no banco de dados IEDB. O cluster selecionado foi o que contém o k-mer “TTQDAYRP”.

No cluster TTQDAYRP, há dez sequências que contêm o seu k-mer e dez que foram agrupadas em um cluster principal, ou seja, todas as sequências iniciais. Então, ao utilizar o Clustal Omega e obter-se uma visualização a partir do Mview, obtivemos que a sequência mais próxima de um epítipo deste cluster é “TTQDAYRPVDP”, conforme [figura 3](#).

Figura 3: Resultado de Alinhamento Múltiplo utilizando Clustal Omega e Mview.

	cov	pid	1 [.] 17
1 6498_2	100.0%	100.0%	-TTQDAYRPVDP SAYKR
2 7104_2	87.5%	100.0%	-TTQDAYRPVDP SAY--
3 18760_2	81.2%	100.0%	-TTQDAYRPVDP SA---
4 6249_2	93.8%	100.0%	-TTQDAYRPVDP SAYK-
5 11541_3	81.2%	92.9%	STTQDAYRPVDP SA---
6 10063_3	68.8%	91.7%	STTQDAYRPVDP-----
7 10061_3	93.8%	93.8%	STTQDAYRPVDP SAYK-
8 12041_3	100.0%	94.1%	STTQDAYRPVDP SAYKR
9 5008_3	87.5%	93.3%	STTQDAYRPVDP SAY--
10 13968_3	75.0%	92.3%	STTQDAYRPVDP SA---
consensus/100%			.TTQDAYRPVDP.....
consensus/90%			.TTQDAYRPVDP.....
consensus/80%			.TTQDAYRPVDP SA...
consensus/70%			.TTQDAYRPVDP SA...

Fonte: Mview, 2022.

A partir da comparação feita da sequência mais próxima de um epítipo, com o resultado do BlastP dos epítipos do IEDB (visto na etapa 3.4 de Validação), foi possível identificar que esse cluster representa um epitopo que já está no banco de dados IEDB de sequência “TQDAYRPVDPSAYKR” e número de identificação “397929”, e portanto é válido.

Partindo de uma outra perspectiva, foi encontrada a sequência consenso a partir da web ferramenta Epitope Cluster Analysis.³⁹ A sequência se configura por “STTQDAYRPVDPSAYKR” como pode ser observado na [tabela 4](#).

Tabela 4: Resultado da Clusterização de Epítomos do Cluster “TTQDAYRP”

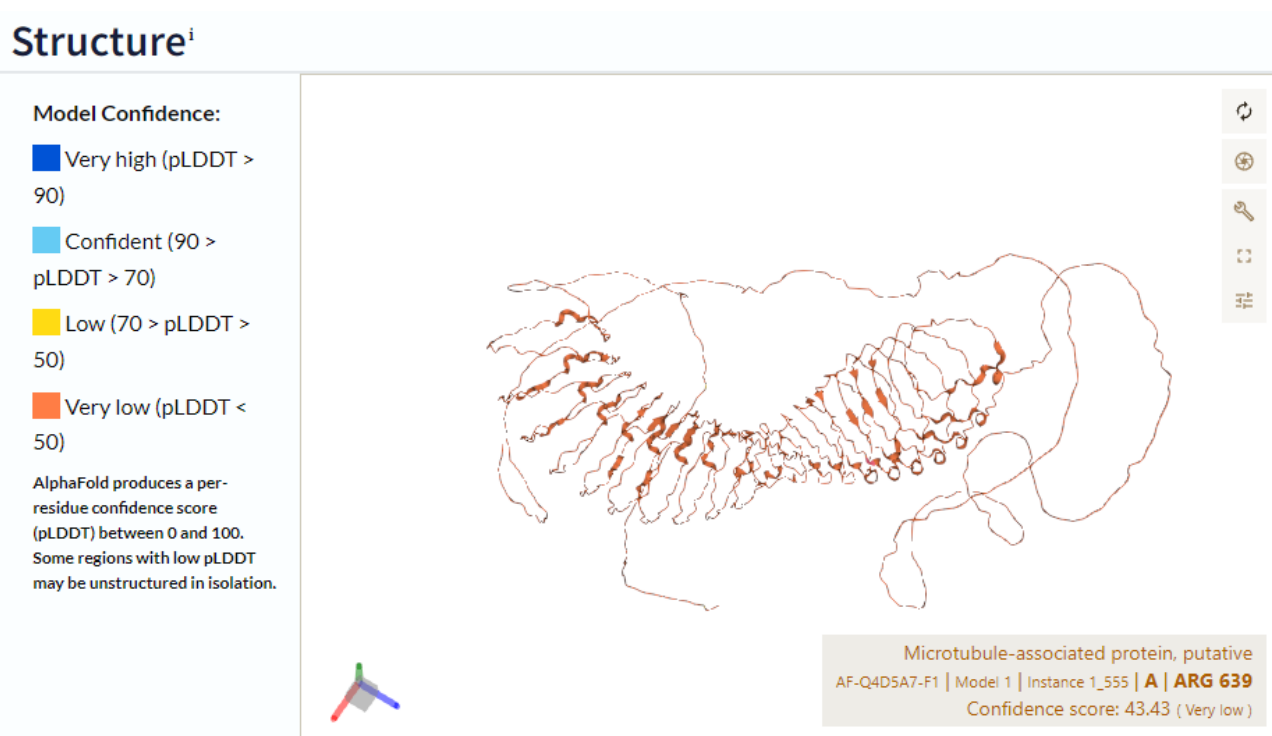
Sequences	Peptide Number	Alignment	Position	Description	Peptide
1.1	Consensus	STTQDAYRPVDPSAYKR	-	-	-
1.1	1	STTQDAYRPVDPSAYKR	1	12041_3 [2 - 52] freq=3	STTQDAYRPVDPSA YKR
1.1	2	STTQDAYRPVDPSAYK-	1	10061_3 [2 - 49] freq=3	STTQDAYRPVDPSA YK
1.1	3	STTQDAYRPVDPSAY--	1	5008_3 [2 - 46] freq=7	STTQDAYRPVDPSA Y
1.1	4	STTQDAYRPVDPSA---	1	11541_3 [2 - 43] freq=3	STTQDAYRPVDPSA
1.1	5	STTQDAYRPVDPS----	1	13968_3 [2 - 40] freq=2	STTQDAYRPVDPS
1.1	6	STTQDAYRPVDP-----	1	10063_3 [2 - 37] freq=3	STTQDAYRPVDP
1.1	7	-TTQDAYRPVDPSAYKR	2	6498_2 [2 - 49] freq=5	TTQDAYRPVDPSAY KR
1.1	8	-TTQDAYRPVDPSAYK-	2	6249_2 [2 - 46] freq=5	TTQDAYRPVDPSAY K
1.1	9	-TTQDAYRPVDPSAY--	2	7104_2 [2 - 43] freq=5	TTQDAYRPVDPSAY
1.1	10	-TTQDAYRPVDPSA---	2	18760_2 [2 - 40] freq=2	TTQDAYRPVDPSA

Fonte: Epitope Clustering, 2022.

A partir da sequência consenso, foi utilizado a ferramenta BlastP³¹ do NCBI, para obter as proteínas que mais se alinham com esta sequência. Houve 45 proteínas que obtiveram 100% de identidade com a sequência, porém, a partir dos dados obtidos a partir do epítipo correspondente encontrado no IEDB, a proteína que melhor representa este possível epítipo foi a 6ª na lista do BlastP, chamada “microtubule-associated protein” de código XP_809567.1, proveniente da cepa do *T. cruzi* CL Brener.

Então a proteína foi encontrada no banco de dados UniProt, onde foi possível observar uma predição estrutural da molécula, como é possível observar na [Figura 4](#).

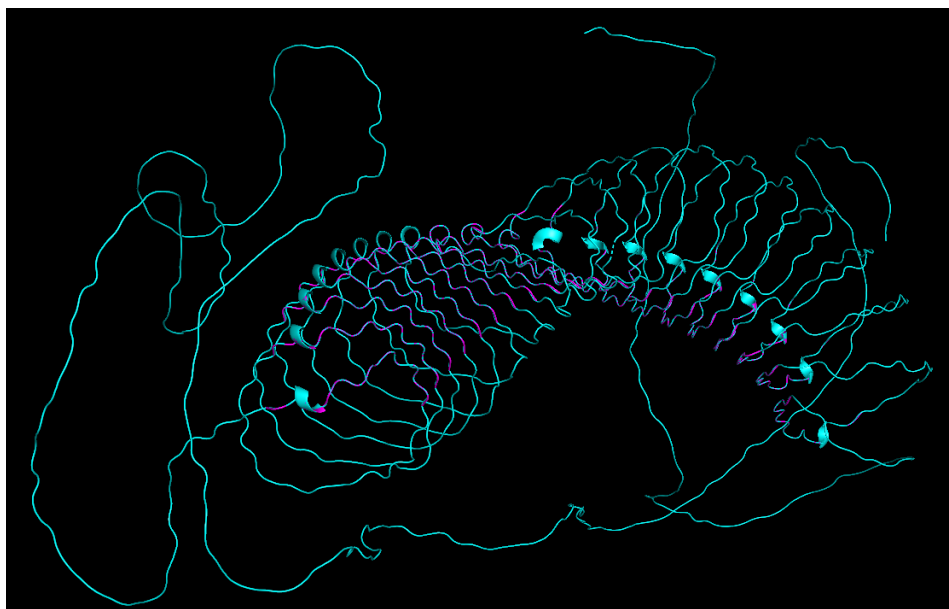
Figura 4: Predição da Estrutura da Proteína “Microtubule Associated Protein”



Fonte: UniProt, 2022. Legenda: A predição foi realizada pelo AlphaFold, e é visualizada no UniProt, além disso contém quatro categorias de confiabilidade do modelo, entre elas com uma muito alta confiabilidade as partes representadas em azul, as confiáveis na cor azul, as pouco confiáveis na cor amarela e pouquíssimo confiáveis na cor laranja.

Então, a proteína foi aberta no visualizador PyMol, e foi colorida a parte representante da sequência consenso “STTQDAYRPVDPSAYKR” que se alinha na proteína, como pode ser visto na [figura 5](#).

Figura 5: Sequência consenso representada na estrutura molecular de sua proteína correspondente.



Fonte: Stéffani Vibanco de Oliveira Neves via PyMol, 2022 . Legenda: A proteína se encontra representada pela cor ciano, enquanto a sequência consenso é representada pela cor magenta.

Portanto, é também possível identificar a sequência consenso contida na sequência completa da proteína, conforme [Figura 6](#).

Figura 6 : Sequência completa da proteína destacando a localização da sequência consenso

```
>tr|Q4D5A7|Q4D5A7_TRYCC Microtubule-associated protein, putative OS=Trypanosoma  
cruzi (strain CL Brener) OX=353153 GN=Tc00.1047053511633.79 PE=4 SV=1
```

```
SVPCRWWSKRMWGRATLIPTTSARRLRTRTGPLIPRRTSAPCRRKSKRMWGRATLIPTTS  
ARRLRTRTGPLIPRRTSVPCRWWSKRMWGRHVDPDHFRSTTQDAYRPVDPSSAYKRALPL  
EEEEEDVGPRHVDPDHFRSTTQDAYRPVDPSSAYKRALPLEEEEEEDVGPRHVDPDHFRSTTQD  
AYRPVDPSSAYKRALPQEEEEEDVGPRHVDPDHFRSTTQDAYRPVDPSSAYKRALPLEEQEDV  
GPRHVDPDHFRSTTQDAYRPVDPSSAYKRALPQEEEEEDVGPRHVDPDHFRSTTQDAYRPVD  
PSAYKRALPQEEEEEDVGPRHVDPDHFRSTTQDAYRPVDPSSAYKRALPQEEEEEDVGPRHVD  
PDHFRSTTQDAYRPVDPSSAYKRALPQEEEEEDVGPRHVDPDHFRSTTQDAYRPVDPSSAYKR  
ALPQEEQEDVGPRHVDPDHFRSTTQDAYRPVDPSSAYKRALPQEEEEEDVGPRHVDPDHFRS  
TTQDAYRPVDPSSAYKRALPQEEEEEDVGPRHVDPDHFRSTTQDAYRPVDPSSAYKRALPQEE  
EEDVGPRHVDPDHFRSTTQDAYRPVDPSSAYKRALPQEEEEEDVGPRHVDPDHFRSTTQDAY  
RPVDPSSAYKRALPQEEEEEDVGPRHVDPDHFRSTTQDAYRPVDPSSAYKRALPQEEEEEDVGP  
RHVDPDHFRSTTQDAYRPVDPSSAYKRALPQEEEEEDVGPRHVDPDHFRSTTQDAYRPVDP  
AYKRALPQEEEEEDVGPRHVDPDHFRSTTQDAYRPVDPSSAYKRALPQEEEEEDVGPRHVDPD  
HFRSTTQDAYRPVDPSSAYKRALPQEEQEDVGPRHVDPDHFRSTTQDAYRPVDPSSAYKRES  
PVVKDVRANVRHAYPDTLRSVSHESYKSVDSAYKRESPVVKDLRANVRHAYPDTLRS  
VSHESYKLLNVASTRDGLSRVCHRISDGKAAQYGESESSFSVNGDRNGTDGASSSCRG  
SARACFGKSSSEVFESNFQTPKGTDDGHFSSKGYFCPCHTDPEMYRSTSHADYKAHHKD  
AYSRPYKPLDRKFPLERRDFLSEYRKNFLRPEPQSLSRPVAASTVTVRHVDPSVYTTN  
QAVFKDHWKKF
```

Fonte: Stéffani Vibanco de Oliveira Neves , 2022. Referência: UniProt

5. CONCLUSÃO

Em síntese, foi empregado nas amostras, obtidas a partir do Phage Display no artigo do Teixeira et al, um pipeline bioinformático em que houve um tratamento inicial do genoma, a partir de conversões, filtro de frequência, alinhamento de nucleotídeos e filtro de identidade; um tratamento do proteoma, a partir da obtenção de ORFs, filtro de frame, alinhamento de proteínas e filtro de identidade; E por fim, uma clusterização e análise de dados, a partir da obtenção de k-mers, então clusterização e identificação de potenciais sequências de epítomos e de uma sequência consenso.

Por fim, é possível observar que a bioinformática é uma inestimável ferramenta para a produção de novas tecnologias voltadas para a saúde humana. Também concluímos que o fluxo percorrido neste projeto possui grande potencial para uma possível identificação de epítomos.

Quanto à lista de epítomos gerados pelo pipeline, podemos ver que há uma grande quantidade de sequências que não foram encontradas no IEDB, mostrando uma possibilidade de identificação de novos epítomos.

No que se refere ao cluster escolhido para aprofundamento da análise, foi possível perceber que apesar do epítomo estar disponível no banco de dados, há poucos ensaios sobre sua função biológica e estrutural. Este pipeline tem a competência de iluminar os possíveis epítomos, chamando atenção da comunidade científica e instigando uma pesquisa aprofundada, individual e possíveis aplicações na área da saúde de cada epítomo.

6. LIMITAÇÕES DESTE TRABALHO

Como em qualquer nova metodologia, a validação completa dos epítomos deve, idealmente, ser confirmado experimentalmente em uma coorte de validação de paciente independente.

Dentro dessa estrutura, os antígenos provisórios apresentados ([Apêndice A](#) e [B](#)) devem ser considerados apenas como candidatos até que sejam inequivocamente comprovados experimentalmente.

7. REFERÊNCIAS

1. Fernandes HJ, Barbosa LO, Machado TS, et al. Meningoencephalitis caused by reactivation of chagas disease in patient without known immunosuppression. *Am J Trop Med Hyg* 2017;96(2):292–4.
2. Perez CJ, Lymbery AJ, Thompson RCA. Reactivation of chagas disease: implications for global health. *Trends Parasitol* 2015;31(11):595–603.
3. Bernstein R. Darwin's illness: Chagas' disease resurgens. *J R Soc Med.* 1984;77:608–609.
4. World-Health-Organization. Neglected tropical diseases. In: World-HealthOrganization, ed. Vol http://www.who.int/neglected_diseases/diseases/en/. Geneva; 2018: Acesso em 20 de Março de 2022.
5. World-Health-Organization. Chagas disease (American trypanosomiasis). Disponível em:

[http://www.who.int/news-room/fact-sheets/detail/chagas-disease-\(american-trypanosomiasis\)](http://www.who.int/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis)). Acesso em 20 de Março de 2022.

6. Garcia M, Woc-Colburn L, Aguilar D, Hotez P, Murray K. Historical Perspectives on the epidemiology of human Chagas disease in Texas and recommendations for enhanced understanding of clinical Chagas disease in the Southern United States. *PLoS Negl Trop Dis*. 2015;9:e0003981.
7. Garcia M, Aguilar D, Gorchakov R, et al. Evidence of autochthonous Chagas disease in southeastern Texas. *Am J Trop Med Hyg*. 2015;92:325–330
8. Bocchi EA. Heart failure in South America. *Curr Cardiol Rev* 2013;9(2):147–56
9. Guarner, Jeannette. "Chagas disease as example of a reemerging parasite." *Seminars in Diagnostic Pathology*. Vol. 36. No. 3. WB Saunders, 2019.
10. Echeverria, L. E., & Morillo, C. A. (2019). American trypanosomiasis (Chagas disease). *Infectious Disease Clinics*, 33(1), 119-134.
11. Yamagata Y, Nakagawa J. Control of Chagas disease. *Adv Parasitol* 2006; 61: 129–65.
12. Pérez-Molina, J. A., & Molina, I. (2018). Chagas disease. *The Lancet*, 391(10115), 82-94.
13. Brenière, S. F., Waleckx, E., & Barnabé, C. (2016). Over six thousand *Trypanosoma cruzi* strains classified into discrete typing units (DTUs): attempt at an inventory. *PLoS neglected tropical diseases*, 10(8), e0004792.
14. Zingales B, Miles MA, Campbell DA, et al. The revised *Trypanosoma cruzi* subspecific nomenclature: rationale, epidemiological relevance and research applications. *Infect Genet Evol* 2012; 12: 240–53.
15. Brenière, S. F., Waleckx, E., & Barnabé, C. (2016). Over six thousand *Trypanosoma cruzi* strains classified into discrete typing units (DTUs): attempt at an inventory. *PLoS neglected tropical diseases*, 10(8), e0004792.
16. De Bona, E., Lidani, K. C. F., Bavia, L., Omidian, Z., Gremski, L. H., Sandri, T. L., & Messias Reason, I. J. D. (2018). Autoimmunity in chronic Chagas disease: a road of multiple pathways to cardiomyopathy?. *Frontiers in immunology*, 1842.
17. Smith, G.P. (2019). Phage display: simple evolution in a petri dish (Nobel Lecture). *Angew. Chem. Int. Ed. Engl.* 58, 14428–14437.

18. Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* 41, e121.
19. Zhang, W., & Reichlin, M. (2005). A peptide DNA surrogate that binds and inhibits anti-dsDNA antibodies. *Clinical Immunology*, 117(3), 214-220.
20. Ellis, S.E., Newlands, G.F., Nisbet, A.J., and Matthews, J.B. (2012). Phage-display library biopanning as a novel approach to identifying nematode vaccine antigens. *Parasite Immunol.* 34, 285–295.
21. Alvarez, P., Leguizamo, M.S., Buscaglia, C.A., Pitcovsky, T.A., and Campetella, O. (2001). Multiple overlapping epitopes in the repetitive unit of the shed acute-phase antigen from *Trypanosoma cruzi* enhance its immunogenic properties. *Infect. Immun.* 69, 7946–7949.
22. POSNER, B.; SMILEY, J.; LEE, I.; BENKOVIC, S. Catalytic antibodies: perusing combinatorial libraries. *Trends in Biochemical Sciences.* v.19, n.4, p.145-150, 1994.
23. PANDE, J.; SZEWCZYK, M. M.; GROVER, A. K. Phage display: concept, innovations, applications and future. *Biotechnology Advances.* v.28, n.6, p.849- 858, 2010.
24. Teixeira, A. A. R., Carnero, L. R., Kuramoto, A., Tang, F. H. F., Gomes, C. H., Pereira, N. B., ... & Giordano, R. J. (2021). A refined genome phage display methodology delineates the human antibody response in patients with Chagas disease. *Isience*, 24(6), 102540.
25. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* 2018 Oct 24. doi: 10.1093/nar/gky1006. [Epub ahead of print] PubMed PMID: 30357391.
26. L.A. Rodriguez-Carnero, A.A.R. Teixeira, F.H.F Tang, A. Kuramoto, M.J.M. Alves, W. Colli, J.C. Setubal, E. Cunha-Neto, R. Pasqualini, W. Arap, R.J. Giordano. Protocol for design, construction, and selection of genome phage (gPhage) display libraries. *STAR Protocols* Volume 2, Issue 4, 100936, 2021.
27. MARTIN, Marcel. Cutadapt removes adapter sequences from high-throughput sequencing reads. **EMBnet.journal**, [S.l.], v. 17, n. 1, p. pp. 10-12, may 2011. ISSN 2226-6089. Disponível em:

- <<https://journal.embnet.org/index.php/embnetjournal/article/view/200>>. Data Acessado em: 12 out. 2022. doi:<https://doi.org/10.14806/ej.17.1.200>.
28. Dias-Neto, E., Nunes, D.N., Giordano, R.J., Sun, J., Botz, G.H., Yang, K., Setubal, J.C., Pasqualini, R., and Arap, W. (2009). Next-generation phage display: integrating and comparing available molecular tools to enable cost-effective high-throughput analysis. *PLoS One* 4, e8338.
 29. Major, Gianluca. `Frequence_Fasta.py`. 2022. Disponível em: <<https://github.com/gianluca-major/pknife>> Acessado em 11 de Nov. de 2022.
 30. National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; (1988) Available from: <https://www.ncbi.nlm.nih.gov/>.
 31. Madden T. The BLAST Sequence Analysis Tool. 2002 Oct 9 [Updated 2003 Aug 13]. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Chapter 16. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK21097/>
 32. Major, Gianluca. `Blast_fmt6_parse.py` 2022. Disponível em: <<https://github.com/gianluca-major/pknife>> Acessado em 11 de Nov. de 2022.
 33. Rice P., Longden I. and Bleasby A. GETORF FROM EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*. 2000 16(6):276-277
 34. Neves, Stéffani. `Fasta_filter_by_frame.py`. 2022. Disponível em: <<https://github.com/svibanco/tcc>> Acessado em 11 de Nov. de 2022.
 35. Neves, Stéffani. `Filter_by_identity.py`. 2022. Disponível em: <<https://github.com/svibanco/tcc>> Acessado em 11 de Nov. de 2022.
 36. Major, Gianluca. `main.py` 2022. Disponível em: <<https://github.com/gianluca-major/pknife>> Acessado em 11 de Nov. de 2022.
 37. Fu L, Niu B, Zhu Z, Wu S, Li W. *CD-HIT: accelerated for clustering the next-generation sequencing data*. *Bioinformatics*. 2012 Dec 1;28(23):3150-2. doi: 10.1093/bioinformatics/bts565. Epub 2012 Oct 11. PMID: 23060610; PMCID: PMC3516142.
 38. Madeira F, Pearce M, Tivey ARN, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Research*. 2022 Apr;gkac240. DOI: 10.1093/nar/gkac240. PMID: 35412617; PMCID: PMC9252731

39. Sandeep Kumar Dhanda, Kerrie Vaughan, Veronique Schulten, Alba Grifoni, Daniela Weiskopf, John Sidney, Bjoern Peters, Alessandro Sette: Development of a novel clustering tool for linear peptide sequences . Immunology (2018) doi:<https://doi.org/10.1111/imm.12984> (ahead of prints) PMID: 30014462
40. Wang Y, Wang Q, Huang H, Huang W, Chen Y, McGarvey PB, Wu CH, Arighi CN, UniProt Consortium. A crowdsourcing open platform for literature curation in UniProt. Plos Biology. 19(12):e3001464 (2021)
41. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
42. Van Rossum, G., & Drake Jr, F. L. (1995). Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam.
43. Cock PA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B and de Hoon MJL (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25, 1422-1423
44. The pandas development team. (2022). pandas-dev/pandas: Pandas (v1.5.1). Zenodo. <https://doi.org/10.5281/zenodo.7223478>
45. Google Drawings. Google. Version 8.1. 2022. Disponível em: <https://docs.google.com/drawings/>

K-mer	Frequency	Unique Sequences	Sequences in the Largest Cluster	Low Frequency Bigger Frequency Medium Size Standard Deviation	Cluster ID	Representative Sequence ID
QAAAGDKP	2013	600	582	14 81 51.0 15.76	0	44523_8
KVAEAEKQ	1045	397	359	20 97 55 14.74	0	27645_8
MRRGQQAQ	945	945	228	21 107 62.0 19.86	2	5433_3
POHRPPPT	747	747	745	25 100 71 16.12	0	41354_5
RGQQAQPH	741	741	174	18 110 57.0 17.7	1	45241_4
PGVFETTG	724	724	624	25 67 48 10.63	1	13669_6
RAQELARE	572	486	339	53 103 76 9.31	1	48266_7
RGQQAQPH	567	567	126	22 104 55.0 20.89	1	45610_3
HSSHTTTR	487	487	130	21 99 79.0 17.63	2	17653_5
RETEHSYG	448	448	237	12 46 14 2.51	3	13889_3
AKAAAAPA	441	84	83	18 80 51 11.50	0	9036_3
RGQQAQPH	407	407	91	42 101 80 12.37	1	33559_5
PAKAPAK	367	198	193	17 102 69.33	0	20739_3
PAAGFGS	355	152	138	12 82 52.0 14.66	0	28444_9
HSSHTTTR	354	354	103	44 89 63 8.68	8	41790_3
QPHSHTT	276	276	74	19 116 103.5 26.36	1	34204_6
RGQQAQPH	241	241	61	34 99 43 18.56	1	2349_4
GNAGGLAP	212	212	212	27 75 51.0 5.91	0	44846_5
RETEHSYG	211	211	211	15 27 23 1.18	0	34666_1
EQERRQLL	191	186	146	28 105 59.0 14.62	0	45628_9
VDPDFHRS	191	162	157	21 82 43 12.49	0	26659_9
HSSHTTTR	184	184	92	22 98 63.0 15.79	1	20148_5
LEQKAEN	170	49	49	36 70 55 6.69	4	1642_2
ANGIRPOP	149	149	60	37 87 65.5 14.44	4	34325_3
TGTLRLD	146	146	103	25 63 40 7.04	0	10847_5
KRQSVNNY	132	132	98	22 55 25.04.78	1	1223_5
RGQQAQPH	130	130	39	17 118 31 25.37	0	23470_5
ORQOPPTM	125	125	44	27 97 60.0 17.44	0	44661_4
HLRQFNPT	116	116	32	27 102 46.0 13.24	0	8703_5
PKSAEPKS	105	13	9	43 81 58 11.71	0	45139_6
RGHHTTHT	92	92	20	36 111 58.5 18.76	1	38362_6
MHRTPHTT	91	91	21	27 64 46 10.21	2	35563_3
EEVPLTGE	87	87	35	22 84 28 10.67	0	17763_6
RGQQAQPH	85	85	15	23 94 27 18.98	1	17875_4
RHLRQFNP	85	85	10	31 69 48.0 10.93	7	17925_3
QPHSHTT	82	82	10	29 90 65.0 18.42	4	3389_5
AAEATKVA	78	52	26	19 93 36.0 19.8	0	40983_7
RHLRQFNP	75	75	9	68 82 82 4.38	2	1656_3
HLRQFNPT	74	74	8	26 85 44.5 16.16	3	8538_4
TKMPEERC	73	73	73	29 59 48 10.16	4	45377_4
QPHSHTT	72	72	9	20 96 65 25.41	0	4547_5
ORQOPPTM	71	71	18	45 92 67.5 13.94	1	14951_3
MHRTPHTT	70	70	11	31 60 38 7.54	2	39388_3
HLRQFNPT	66	66	8	30 82 51.0 16.14	3	6132_3
DNMRRGQQ	65	65	16	40 46 44.5 1.94	9	34172_4
EDLQRLE	62	42	32	34 70 54.0 7.09	0	29003_6
ARRLAEEA	61	15	6	32 72 53.5 13.43	0	7841_2
RGHHTTHT	60	60	7	22 116 66 28.46	0	36531_6
MHRTPHTT	59	59	9	21 79 36 19.81	0	26227_4
HLRQFNPT	58	58	8	31 79 56.5 15.36	4	1942_5
KARHWLPL	58	58	57	20 24 22 0.53	0	39008_3
RAPEQVPH	57	57	29	35 86 53 14.73	0	45725_4
RGQQAQPH	56	56	7	52 77 64 9.79	8	9692_5
SLEVERKK	55	55	55	25 99 67 11.62	0	42919_6
HHRPPTSH	55	55	51	19 56 38 9.36	0	21891_3

YYTNPVRT	43	43	4	30 83 33.5 22.06	0	43062_5	LTTSKQKFYTHFFNYISVPIPTYPYTLNLTNFTNLNQHYKLSISSLYTNPVRTPPPRDNQPFYRNNVRGRCMIFRPQI
MHRTPHTT	42	42	5	26 56 47 10.35	1	24636_1	NHMRKQQPQMHRTPHTTHDTHSNPRQOAHTRQWRKREKQTGAHKAPPHAHTAR
HLRQFNPT	41	41	3	49 69 5 18.99	6	44584_4	RGFGPPHNKRTHRQOQSTNEREREETHRHRLQFNPTANGIRPQPRSQHTAQTEAQTORATRRSSDLS
TPTPIEPH	41	41	7	15 49 27 11.04	2	13518_3	LRSIPRHHQPNHNYHYQTNAYITPTPIEHLPVTTPKIGQIMYGGDA
PSHAEHST	41	41	38	29 103 61.0 14.63	0	23639_9	RPDRTTSAPSPLCSSLRLVSSCSPKFLSRSSSPSRSDISLPSPEATHNTHRAKRKYHTSVYRPPWPSHAEHSTGAADRHFPAHRHHTHAQQITGTH
SHITRRHP	41	41	5	40 84 81 17.59	1	12705_4	IGEPECKGMERKGEGRSTADTARKRADNMRRGOQAQPHSSHTTRHRPRNEGATORQOPPTMOEKTSTTISPSFPOQNGKNT
RGHHTTHT	40	40	5	29 108 34 30.03	0	46783_5	EKGMMQOGEGRSAADRARESRQHEKRAAVPAKVYTHETSTOQRNTTPAATHHAGGEDIHHHLPOLIPTANEKWQKHRGHHTHTERAPTTPATRGNKISSHPALN
GFGQQAGG	40	24	18	24 84 43.5 10.75	0	42172_3	AAGRTAGAGFGQNTTAGGFGGQTAATGGFGQQAGGFGQQAGGFGQQTGGFGQAAGRTAGGFG
NGIRPQPR	40	40	13	38 93 57 14.53	0	41181_4	QGRNKRQAGKAPPRASPWPATPHQPSHIHTPSQCCQSAGERIWSSTQQTHLPAINHRRGQGOQEAHRLRLQLNPTDNGIRPQPRSQHHTHTKQ
FPNDISP	39	39	32	28 100 65.0 18.42	0	21144_8	HSPAQPSSESGPVIFKQSSDVEIPTSAGVGMAEESPGSGALAPASSQTQNASHELLGTEMPVSGEHFFPNIDSLPMLGQVDTADEESPRIQNTDQ
YYTNPVRT	39	39	4	31 51 36.5 8.07	4	20502_3	ILIDNHIHKPTVSLPLYTYTNPVRTPPPRDNPKQRNNVRGRCMNFPQI
HLRQFNPT	38	38	3	45 59 59 6.6	9	31858_3	VTHQRRGRENHRHLQFNPTANGIRPQPRSQHTTTEQTTQTORATRGSDSSSTHDAAP
AATEKRQS	38	38	31	20 71 39 13.02	0	1500_6	YPATTEGTAQSTTSGSQEQEAPSTSEEPSFEEHQSTGKTTEDARTPDAAATEKRQSVNNYTTTPGDS
MHRTPHTT	37	37	5	30 46 34 5.46	7	18844_3	QRANTOEKRGQQPRMHRTPHTHTHNSNPRQOAHTHSPAATPTTHR
HTHTQPCR	37	37	29	25 63 40 9.14	1	17607_3	EGSSOACTVVRTPTHTPIHGSRHHTQPCRNNTQMEEGEANRSGTQGSTACTYRESGPSC
SHITRRHP	36	36	5	41 72 49 10.54	5	19427_4	GKRMQOGEERRSTADTARERADKMRRQQAQPHSSHTTRHPPSPSPSFLPGNGKTPRPPHNAHGASTHH
RGHHTTHT	35	35	5	27 67 32 14.99	7	27487_4	HHHLPOLIPAAPGKWPKPRGHHTTTEQAPTTGRSGSRASSIHDAQNFLLPSRPQWNTPEPNSQSH
THRHRLQF	35	35	5	30 65 35 12.53	4	5643_3	VSPRVRGFGPPHNKRTHRQOQSTNGOGREATHRHRLQFNPTANGIRPOPAANKHNAQSNHLKS
EARRLAE	35	9	5	40 64 43 9.11	1	34388_2	ARRLAEAEAEARRLAEAEARRLAEAEARRLAEAEARRLAEAEARRLAEAEAEARRLAE
RWFNLTRG	35	35	35	30 49 39 4.49	0	37626_5	LRRGSSSRWFNLTRGSRSPSCSGIPLRCEGSPGRSGVSGSGESI
APYAAHHA	35	35	11	32 65 49 8.71	1	27097_3	TGETRGAADTHERTSHHEERAAARHAPYAAHHAQHTLQSTAEQTHHTHTHTSPAATTHRWKREK
YYTNPVRT	35	35	3	37 69 49 13.2	0	19414_5	NLSTSNSTNTSFHLNKSCLYTRYNNKPLNLTIRYCTLYTNPVRTPPPRDNPHFQNNVRGRCMK
HLRQFNPT	34	34	3	37 54 52 7.59	12	6493_3	DEKIRPSTQRRHTLPADIRGEEQEAQRHRLQFNPTANGIRPOPAANKHNAHQ
TPTPIEPH	34	34	5	18 37 20 7.07	6	712_2	TNSYLHLSALYITPTPIEHLPVIAKSVIEMYGDDA
PPSPPAH	34	34	4	35 72 72.0 16.02	6	37349_3	RRTTHPPSPPAHLYSPREMAKTPRSPHQPHGASTHHPSTRGNKISCHPVLNTRGTPQRSHTRCTTPTGKE
RGQQAQPH	34	34	5	38 61 55 9.87	8	6291_5	QQOGEWGQADTVSEADNRRGOQAQPHTSHTTRHRPKDGTORQOPPTMQEEKISTISS
DSSAHSTP	34	12	10	43 73 56.0 9.72	0	47656_8	DSSAHSTPTPVDSSAHSTPSTPADSSAHSTPSTPADSSAHGTPSTPADSSAHSTPSTPADSSAHSTPSTPAD
MHRTPHTT	32	32	5	23 41 27 6.22	10	32579_4	TGETRGAADTQRAHHMRRRQOPAMHRTPTHTDTHSNPRQH
YYTNPVRT	32	32	3	25 48 35 9.42	4	22265_5	NYISNNHILYAPLYTNPVRTPPPRDNPLKQNNVRGRCMKFRPKI
SKQSPKEP	32	32	5	50 96 68 15.55	0	25931_4	QRCQSAGERIWSRQONTSPVITHRRGGTTHRHCLQFNPTTNGIQOPLTQRTTQSKQSPKEHAKAIPHPSMTRPQHQFSQHPQWNLMPDSKRS
NMRRGQQA	32	32	13	16 81 26 23.45	2	13411_4	KRERDGPRIQESADNMRRGOQAQPHSPHTTRRRPRKDGTORQOPPTMQEETSTTISPSFPRPHOPHNEQADPHPT
HHSLRPHG	31	31	19	32 62 48 7.04	0	12044_4	AHPLLQGLPGACQVHPVTOORREEQKPGVDAAGLRDRDEGRHLSLRPHGGRGPGRAEVGQ
HVELQRRN	31	31	11	24 58 30 10.31	0	34881_3	INGRPAANTIKEVQPHVGAHWSTAATLRSPHVELQRRNHKSHEVPSNQPCGE
QOQQOQQQ	31	13	4	33 46 41.5 5.68	5	3442_4	VFHRFMHGDSEMRDGVQRORPOQOQQOQQHCRCFEEDVEEEG
RGHHTTHT	30	30	4	51 65 55.5 5.21	8	28141_1	PPSPPAHSRPREMPKPRGHHTTTEQANPHPSITRHKITCPALNGNTRSPQRSQMRCTTPAGK
PRPPHNT	30	30	11	46 55 50 3.32	5	27213_3	TRREGSRPSSAVHTPRDGRTHHHTPLISTTPGKWKPAPRPPHNTHGAESPVGN
HLRQFNPT	30	30	3	38 49 38 5.19	13	6653_4	YTPPAVTHNAREVTHRHRLQFNPTANTQRKWSQSPKEHEADPHPS
LTVRHGAH	30	30	8	39 71 48.5 11.86	0	29849_4	ALSTHTAATKETGGSNSTRQSTLSRTHGLAERLTVRHGAHPLSSECTQOASAQSPSCHRNGNKQSSSQ
TAHNRVHP	30	30	30	19 80 34.0 11.87	0	36679_3	MAARGRTPRESIAPPPSPFPQSPARTAHNRHHPHSPRRPQRPNSQPKNRGNHTPATKHATQHRVTVTPPSEECHV
TPTPIEPH	29	29	4	21 47 29.5 9.47	4	39868_4	IHHDLKHTVNLINYTHAYITPTPIEHLPVTTDPFGEIMYGDDA
YYTNPVRT	29	29	3	33 39 35 2.49	8	26952_4	SQTYDLNKKLKSQIQORNYSLYTNPVRTPPPRDNPHF
PAAGFGS	29	14	14	26 76 37.5 14.64	0	36001_3	HTSAPAAAGFGGSAHSTPAVGGFGSATTTSPAAGFGGSAHSTSPAAGFGGSAHSTSPAAGFGGFSATTTSPA
PPSPPAH	29	29	3	89 97 93 3.27	0	35758_5	TTGQHPNRNANASSHPCTRRYPPSPPAHSHSQREMPKARGHHTHTERAPTTITRASGTSIHDAAKLPAIPPSMGHGAQLNAVTRGAPHQ
RAPPQVK	28	28	17	33 86 57 13.82	0	18002_7	SQEESSQAASPVKPSPEEIGKKSQVTKNVFLYNRLNSTERTAKDRKPKRAPEQPVQKAPKPVAPAAPAGPREVPAALGR
PQATOQRG	28	28	28	34 88 63.0 12.76	0	11556_4	PKRTEETRRSQTSPTASQARSLWLQSSOKETHTRTTPSTSPERSLPQATOQRGSGCGSLTRPAPQSPRTRQREOTRMGSSADTA
TANGIRPQ	28	28	9	29 88 43 19.73	2	23768_4	ALGTNSTPALKHTHRSASPRVRGFGPPHNKRTHRQOQSTNGOGREATHRHRLHFNPTANGIRPOLAAGTORTRSEQSPKEPHAEAD
HHLPQLP	28	28	11	36 95 54 16.88	0	44503_4	NEKNGTRGRGTEHSGYERERRQHEERAAQPAQSTHETASTORRHTTPAATHHAGGEDIHHHLPOLIPAAPGKWQKPRGHHTTYTEQANPHPS
MHRTPHTT	27	27	4	46 53 50.0 2.86	2	24828_3	EGHGEQQRREVROTHTORANHAGRGQOPGMHRTPTHTTQOSLKAKONNEKISQ
VDAAGEGQ	27	27	25	37 72 55 7.42	0	29473_5	ERLFDVGGLRGGVPFDGEPGRVWDAAGEGRARWRLEGGDLRRWRDRWRDRCEWHAELCAQPSAAAAV
YGPLRPTG	27	27	26	44 91 48.0 10.83	0	10980_6	FSKSHVLRTTGGDAAAGTPQEVSGVSTSEPTDQGMPEQDYGLRPTGMWNVEEVVDYKNSVDVFRIDDEVESEVIELSQPDVAVPVE
HLRQFNPT	27	27	2	69 88 78.5 9.5	2	28949_5	EETNDRREKHRQPHPDORQLHTCPHTKHTPSCRQSPGERICSSQTQHTTPPAVTHNARGEETHSHHRLQFNPTANGIRPOPAANNI
QREAEERA	27	11	8	25 65 54.5 12.22	0	30426_2	EERAQREAEERAQREAEKRAQREAEKRAQREAEERAQREAEKRAQREAEERAQREAEERAQREAE
KGQTSLED	27	27	7	44 99 68 13.73	0	12764_5	TEGDKPPVNRLYGRPSSLTVSPKKSATKKGQTSLEDHRSPGTRASRSNRPVPTTKRPTATRNTTKEGKSRRPGDLSLPGINADFAERIESEIIRSP
IRNEGSVS	27	27	26	48 131 87.5 29.07	0	2061_7	GTRNDGYGNASNRRDRVEWDDRRDGMRTDMRLIRNEGSVSRNRPERANRNGRYDERNERNERRVERHDMRGMRGENDNVRRARDERGVLGEPAPQSTRRLRSREQPSTAASKSENG
QOQQOQQQ	27	9	3	57 61 58 1.7	2	28495_3	NIKQOQQOQQOQOPLQKKKKRKLRLRLPKRWGRNKHPTHHTHAKHTHTHTQEREGE
HRTPHTT	26	26	7	27 56 37 9.2	0	13638_4	DMESNRGDERGRHTQRAHRMRRGOQGMHRTPTHTGHTSNPRQHAHSPAATPTH
GTORQOPP	26	26	4	55 91 62.5 14.31	0	29199_4	QAQPHSSHTTRRLRKEGTORQOPPTMOEKTSTTIFPSFPOQSGSKTPRPPHNAHGAEPPSMTRHONFLPSRTQWETEPNSAQSHAV
YYTNPVRT	26	26	3	21 34 27 5.31	11	8843_3	SLISTQLLYTNPVRTPPPRKFRQIDKIMYGDDA
SKQSPKEP	26	26	4	47 65 56.0 6.71	5	11612_3	HNKRTHCQOSTTVDDKKRHTATASSIPRTTGYGHSAPVNTQRTPSKQSPKEPHAEAIHPMSMT
PPSPPAH	26	26	3	72 94 76 9.57	0	9491_3	DTAREREQTTRGEGSRPSHTVHTPRNGIHERKENASSHPCCRRRRHPPSPPAHSHSQREMAERNRGHHTHTERAPTTPTVTRASEPSSIHAA
RGQQAQPH	26	26	4	41 91 68.0 23.57	0	3868_4	EEKECNKRERDQAQRIQERADKTRRGQQAQPHSLHTTRHPRNEGATORQOPPTMOEKTSTTISQLOPQTGTGHSRPSHNAHGASEPS
TPTPIEPH	25	25	3	18 52 22 15.17	1	43964_3	TOSNLIVTSQHSPSNITQSNINTVTINHLNLCYITPTPIEHLPENANSNKT
THRHRLQF	25	25	4	54 86 74.5 13.01	0	42220_5	KRHPHSHKHTTPVTHQOGEVYTHRHRLQFPTDNGIRPQPRSQHTTQEQAISQATRRSDSSIRDEAPHSMRSHGAIHFLNRTH
HLRQFNPT	25	25	2	72 73 72.5 0.5	3	4982_3	TQQTHTPAATHRGRGEETQRHRLQFNPTANGIRPQPRSQHTAQTEAQTORATHAEAIHPSTTQRPSPG
THTHSPAA	25	25	7	28 74 47 14.7	0	33225_5	DLHSEHGEGQGRREVQTHTEGTPHEERVAARHAPYAAHHAQHTLHSTAAGTHTHSPAATPTHWKREKQTGA
VQOARRTG	25	25	19	12 58 16 12.61	0	42724_1	LRVSAPRRKENKTRGKEMCAPHRPSSHAHTRSLPQRAVRRTGSSCQVQARRTGPSS
TTTTTTTT	25	12	3	41 54 49 5.35	0	27931_3	TKKPTTTTTTKPTPTTTTTTTTAPAEPSITTTPTNTTTTRAPSSIRRIDG
PKSAEPKS	24	4	4	48 77 61.5 11.9	0	47942_5	PFTEKKPAKASTATSPSEVHTVTPVATEPKSAEPKSAEPKSAVPKSAEPKSAEPKSAEPKSAEPKSAEPKSA
FHPSTPLP	24	24	17	27 73 42 12.14	0	10269_5	EASATRDVANDGVSTLPHQTDILMAGDVSSAARRSSAAHRPPHNGAARPRFHPSTPLPSRTSPLDARLPMEE
ERREARER	24	14	14	35 47 42.0 3.36	0	40736_4	REERRGVEERREARERARHEAKERREARERARLQKLELTVTTIKD
GRKCKPTL	24	24	24	21 81.51.5 15.94	0	36331_5	KQRKKGIAESNSLPDVRCEGGRCKPTLRPHTPHAFARHPSPSRSSSGSSSLPYAEPASTFAPPSPASRAVPSPFPEEVKL
WRWAWHSH	24	24	13	17 46 29 9.16	0	33917_4	SQWRWAWHSHGGAVGAPCSRATATPYGAEAAHRPSRNSATCDQKH
SHITRRHP	24	24	4	52 83 70.5 11.51	1	25332_2	ADNMRRGOQAQKSHSHTTRHPRKEGTORQOPPCRRRRHPPSPPAHRSRPREMAKTPRPPHNTHGASKPSIIHAAQKFP
QQTGTHV	23	23	23	31 63 40 7.81	0	42302_4	SVYRPPWPSHAEHSTGAADRHFPAHRHKKHAQITGTHVTRTPNKGQOROROTEQORTKPTTS
YYTNPVRT	23	23	2	36 61 48.5 12.5	1	47345_4	RLKYVSHHHYQPYQNNHKLISYTLTYTNPVRTPPPRDNPHFQNNVRGRCMNFPQI
HLRQFNPT	23	23	2	30 59 44.5 14.5	4	2508_3	VTHQRRGRENHRHLQFNPTANGIQOPQRSQHTKHSQSPKEPHAEAIHPSTTRWQN

TPTPIEPH	22	22	3	25 48 319.74	1	19664_3	KSINKQNTKDFITLHPPTTAYITPTPIEPHLPVKSANFQKIMYGDA
SPPESTFA	22	22	15	22 22 22 0.0	1	14447_3	FLTALRRIDASPPESTFAAPR
LTVRHGAH	22	22	5	24 61 43 13.02	0	43962_4	TKETDTSNRRRALLSSPADWRPTLTVRHGAHPLPSECTQQAGAQSPSCHRGKNKQSSS
HESFLSLE	22	22	15	42 102 85 16.48	0	21607_7	NQFGQPYMSSNSKTADVSSIFSTEHGKGDPHKEATPTLTEIQPPQOLRQOKRHLHESFLSLEKTGTVEVERLOVSSPRRREKTEPRRPAQEGSPVFLPPEVSA
RWHTLRSE	22	22	22	23 57 40.5 8.17	0	12570_3	LDRAIAAPDAEDGRRLDLRREALREASKSRWHTLRSELEGTODGNRQHVVRPCLSP
NPRQOAH	22	22	6	30 44 35.0 5.25	6	36711_3	PGMHRTPTHTDTHSNPRQOAHTHTHSPAATPTHRWVRWEKOTGA
QALGRHSH	21	21	17	26 41 32 3.79	1	33861_3	VSTLTPRHAQGRSQAOLGRHSHQGRADGYROOGLSTVTDVA
VYEKPHQ	21	21	21	40 99 65 12.12	0	19240_7	KGESGNKIGPAAHATVATHSDNKNREHKTKAVYEKPHQKKGPEKPKSKNNKNKDGAKSAENLNESLPLPLSPRSNGVVEEASVAPVNVNTPIVEDL
YYTNPNRT	21	21	2	32 46 39.0 7.0	4	28622_3	YNTKINKYHKPTVYYTNPNRTPPPKIHGFGRNVNRCRCMNFQJ
THRHRLQF	21	21	3	25 72 66 20.89	2	24074_4	TPSHSAHTAKRHPHSHKTTSVTHQOGEVTHRHRLQFSPTDNGIRPQPSQHTHTTEQSKPEHAEAIHPHS
SKQSPKEP	21	21	4	45 62 54.0 7.79	5	44733_3	IASSSPIRPTGYGHSAPANTKHSQSKPEPLAEIHPMSMTRRONQFPQPKWNLMPSDKRS
PPPSPPAH	21	21	3	52 77 63 10.23	1	21141_4	ADRWDRKRMRQREGERRSAADTAREADKMRRGQQAQPHSSHITRRHPPSPPAHSCCPRMAKPRGHHTHTTEQAP
ELQRRNHH	21	21	12	18 56 36.0 9.75	0	34114_3	EETQPHVGAHWSTAAATLYSPPHVQLRRNHHKSHGEVPSQHPGCEESHSHRPTA
SAIGPAPE	21	21	21	32 68 56 8.57	0	29131_4	ATAEFAARAPNSAIGPAPEERSRRMEASGRRHHPAGAGKSEKREEEAPPRGGDCGEPLDVANLSPLLT
TRAPSRLR	21	21	5	27 38 35 3.9	3	48354_3	AKTTTTTTTAAPEASTTTTEASTTTTTRAPSRLREIDG
PGROLHTS	20	20	9	19 82 35 16.75	0	46381_3	KDRQTGGEMNGRRQDALINMTREHAATPLRNTGREGEKSKQHNTHGRNKLREKHRHSPRPGRQLHTSPHTHIAVPSVPG
APYAAHHA	20	20	8	45 61 49.5 6.34	0	45153_3	HEERAAATHAPYAAHHAQHTLQSTAGHTHSPAATPTRRWNGEKQTAAGHAKPPAHITAR
PTASTHRE	20	20	10	33 48 39.5 4.3	2	23265_4	TRTQRMHPQRLPTASTHREVRALEWEGRSRORTQKEKHAEHPSTNS
HLRQFNPT	20	20	2	50 57 53.5 3.5	5	47004_3	QQLWRTACVNCETAPTQOHTTPPAATHNGOEATHRHRLQFNPTANGIRPQPAAGTOR
AVHHHRPS	20	20	11	27 70 40 11.62	0	39947_4	HLAPRHAQHSALQRNHARTVRKRVNATEHSRVORDATDKLOHSAVHHHRPSTSHHRIEGNAGHNQHGPH
KTRFGTRW	20	20	19	22 83 48 16.23	0	16114_3	GRSEHPHVLKSHRRSQPSLAASGMPSRDGHKDVASNKTRFGRTRWANDDRHCCMGCPSAHLKSHFYASSSSSSSINTPVH
CSRVSRPP	19	19	6	24 63 40.5 15.62	0	10116_4	PSNARTRQSKNKNKREKNDPKNRQHLLAWRASCSCRVSRRPQAQKRRPHTPSNGTVIPTAGTHAA
LPVPSTGE	19	19	12	31 54 42.5 7.03	1	18188_3	NGAIEGHAGESDVHAKSSSVSSAESRPTGKPSLPVSPGTGESPTTAGDTRNST
VKAERK GK	19	19	11	37 65 53 7.31	0	11676_3	FGPVKAERKGDAAAAPARAEEKPKAAAAAGAAEEDEAPREKKKPNLDELPSPFVLDFAKRE
TPTPIEPH	19	19	3	24 34 27 4.19	3	8060_2	KNSKMYRPSYITPTPIEPHPEKSLFGQISSWF
AQHTHGRN	19	19	18	29 79 47.5 13.28	0	25067_6	QPAQVPRTHHTDPLSHISTHKGEDISKIPPHRNEGPWAQHTHGRNKRDGTHPSTRRSLLASPTDRHHASQQAVA
RHHPHSWG	19	19	19	17 48 24 8.36	0	11424_3	ILFVPSDGGTKLAAAEVATGTHAHHRHPHSWGGRGTGTPSHQGYAQI
NKSEHETR	19	19	11	33 84 51 15.32	0	46010_6	DSRVNKGQSRVTVGLVPTGHTSGGASVTAHTLTPGATPPALTAVAPWADPFTNKSHEHTRYPNRKHPPRGEGSGVGEPKQA
PAHFHNPR	19	19	9	33 83 58 12.62	0	39812_5	RNRAQRIQERVDNTRGQQAQPHSGHTTRRHPPPPSPHAFHNPREVAKTPRPHMHTHGASEPSVSNDAATKFTLPSSPHWN
YYTNPNRT	19	19	2	30 32 31.0 1.0	9	45840_3	TLPRLLYYTNPNRTPPRDNPNQPKIMYGDA
EEGEANRS	19	19	5	38 57 40 7.14	1	33209_2	RTHTPIHGSRHTHTHOMEEGEANRSQTGACTYGESGSPCLHRSRRARTHRE
YQHPPHRH	19	19	16	21 47 35.5 6.63	0	17055_3	VQGEAPATRLYQHPPHRPPAPSPPRHTSDARRPSEDAPAVAROR
RGHHTHTT	18	18	3	46 72 70 11.81	2	48259_2	ARESQRHEERAAGPAQFHTHETAPTITTPRSFQLQGNGRKPRGHHTHTHQAQNPQLVMARPNOSHCPV
FLDQKPEG	18	18	17	22 72 65 10.89	0	23284_5	AQELAREKLDRAFLDQKPEGVLPURELPDDSDVFAMEQEROLLEROPRRNAREIAALEESMNARAQEL
HLRQFNPT	18	18	2	26 27 26.5 0.5	15	5767_3	HTNARGGETHRHRLQFNPTTNGTRLQ
PPPSPPAH	18	18	3	32 64 45 13.14	3	7685_3	HATKAHNASSHPPCRRRHPPPSPPAHRYSPREMAKTPRPPYQHPAPANPHLMTRHKIKCHPA
TTTTTTTT	18	9	2	37 44 40.5 3.5	1	26042_3	TTTTTTKPTTTTTTTTQTPTSTTTTEVPTVSTTRAPSRLREIDG
QQQQQQQQ	18	6	3	44 51 44 3.3	3	20956_4	VAALARSQSGSGRGYQOREQQQQQQQQRFGGPRDGRPLRGRPPNQFQG
TPAATHHA	18	18	6	55 86 72.0 11.6	1	21435_4	RQHEERAAGPAAQFTHETASTQRRHTTPAATHHAGEEDIHHLQLPIAAPGKWQKHGRHTHAGASEPSSIHDAAPKFAIPPS
ASPSGRPS	17	17	10	29 42 39.0 4.68	1	24371_1	RTHSHRCSAQDQWPRSPRTAPQGAQSSQCAPSGRPSASCS
HTHTTEQA	17	17	6	68 82 75.0 4.04	0	42143_3	ESKKKONTEHGNKRGAGKAPPRHTAGNSPALSHTHSHRHLQFSPTGNGVROPQPSQHTHTTEQAOTQRATHGGDSSSTHD
DEAWRRIQ	17	17	2	26 38 26 2.82	0	10549_3	AEAMKRFATQKEKSEKFIQENLDRODEAWRRIQELRV
LTVRHGAH	17	17	5	26 44 29 6.69	2	2951_4	HSRSWQWRPTLTVRHGAHPLPSERTHGOAQSPSHRNGETQTI
RHPRATRH	17	17	9	25 71 40 12.18	0	27447_4	KKHTAEAAATKLPCCADATARKGAPATESIAQETEKSPTRHPRATRWQSGRCGLNATTTNSAATPECT
SKQSPKEP	17	17	4	40 58 49.5 6.46	5	5141_3	TTSVVTHQOGEVTHRHRLQLSPTDNGIRPQPSQOTKRRPSQSKPEHAEAIHPHS
EVPLTGE	17	17	12	20 39 32.5 5.57	2	44900_4	SLGEEVPLTGEAPLGLVHFCGACGEDAGOKTVTVKN
QPHSSHIT	17	17	4	20 88 43.5 29.69	1	9826_4	NMRRGQQAQPHSSHTTRHPRNEGTORQOPPTMOEKKTSTTSPSSFPQPOGNGKNTAATTHTHEQANPHPSMTRHQNFLPSRPQWDT
HAMRPLTK	17	17	16	25 43 29.0 4.64	0	2955_5	NLYFPHRTAVPHAMRPLTKKPRLHDSSNEASOGSTWNSLPPS
MHRTPHIT	17	17	4	41 52 50.5 4.5	1	9201_3	HMRRGQQGMHRTPHITHTNTHSNPRQHTHTHSPAATPTHMEEGEANRSQTQ
YYTNPNRT	17	17	2	20 30 25.0 5.0	10	24336_1	GLVTDPSPLYTYTNPNRTPPPRKIAVFGPN
TPTPIEPH	16	16	3	25 30 27 2.05	3	2887_3	ISARPYITPTPIEPHLPVTTNPFTEIMYGG
LQHYHRGG	16	16	16	37 98 68.0 16.29	0	40585_3	NRGYSGGFHDRRHFDYRDGPYSGNGDGVPHLQHYHRGGSGGGRSRGGGSDVPQPTSHRSNENHOEKKRRWTQNAKDTAGATTDGEKEQQRKPGN
AGWRRQGS	16	16	10	29 45 41.0 4.25	1	38469_5	TAREKERYAENAMPLHNAAGWRRQGESLYPTLRRGQRSDDPVCS
THRHRLQF	16	16	2	74 92 83.0 9.0	0	30246_3	HAESGTQENPEGTARSTTAHTSPHTQTHAHSQCCQTARGIRPSTQKTRAPPATFTHNGRGKETHRHRLQFSLVANGIRPRPSQHTTLTVP
HLRQFNPT	16	16	1	92 92 92 0.0	0	23940_5	KNTSLGKSATALTLAGNSTPALHTLTPSQCCQASAGERIRPSTQOHTTPAATRHGRGEETQRHRLQFNPTANGIRPQPSQHTAHTEQA
PPHNAHGA	16	16	3	24 49 45 10.96	8	21813_1	RRTHPPPSPPAHLYSPREMAKTPRPPHNAHGASTHTHTSHTRQRTLHP
SSDIEPF	16	16	5	51 76 55 8.91	1	45031_7	SPRIQHSQAQTSSESGPVISQSSDVIEPTFADVMAKEESPGSGALAPASQTNQAGSHELLGTEMPVSGDH
HDTHRCRP	16	16	16	19 30 26.0 3.06	0	46108_3	TPMQARWCSLHDTHRCRPAWLHGRSTVTV
SSEGRCSR	16	16	11	26 40 31 4.72	0	43089_3	TWTSRQPSLAPASRSSAATSSEGRCSRWSVCSRTPRTVTSV
RGHHTHTT	15	15	3	45 71 64 10.98	2	3487_4	EGSRPSRTAHTPRDDIHHLPLISTTPGKWPKRPGHHTHTTEQAQPTRGRSGSRASSIHDAQNFQPSRP
RSSAQOGV	15	15	15	12 43 16 12.85	0	1322_1	PRTPSHHRSQAQGWPRWSRTAPQAQSSPCASPSGRPSASCS
GSQVQRG	15	15	15	46 80 62 9.96	0	43818_3	RSQSSLERNINTPAQAPRRTOQEFRSGSQGVRRGASHEORQPOSSATPSLPGRRVEAQQLPHAGFRQGLTRPNKNPYAIA
HRHRLQFN	15	15	3	58 85 81 11.9	0	2292_4	LNLGNSTPALHTPSQCRQASAGERIRPSTQOHTQPAVTHQGRGRGNHRHRLQFNPMANGIRPQPSPHTKHSGKSPKEPHAE
EARRLAEE	15	4	4	47 72 62.0 9.36	0	35245_3	IRKTTNSKTNSNTKEEANILVKEAEARRLAEAEARRLAEAEARRLAEAEARRLAEAE
ECWHAQPA	15	15	6	36 51 41.5 5.27	1	16897_4	LLYSDHSRHTVLTCEWHAQPAHTSSAQSGIRQCGCGEQQRGTHPRDDA
TFGPSHAK	15	15	12	27 44 35.0 5.45	0	40702_3	TRTRPHITPVKTFGPSHAKDWSGRTAGKTTTAAAMSSPHTPHWR
HLPHHHPF	15	15	10	37 59 45.5 6.93	4	41381_4	TLLTQSDQPTVPAEAAAAAPSRRRKHQHHLPHHHFPLVPSLHSSQKSSVLAHVGYWR
THSNPROQ	15	15	5	30 56 44 8.33	0	26490_3	QRANTQEKRGQDQGMHRTPTHTNTHSNPRQQAHTHALPQQQHTDGGGRSKOER
DEYSPEKA	15	15	15	38 75 49 10.74	0	14504_5	RTPDHPSDRHEGHQRHPQPKHREAPFQWRHQTHTTVEYSPEKATLKOPQTLPLLARHPVHPHTHSERQD
TTTTTTTT	15	7	2	35 36 35.5 0.5	2	1084_3	IQRGQWQLDRGRYHSPPPPPPTTTTTTTTSHSD
YYTNPNRT	15	15	1	66 66 66 0.0	0	37653_6	QTHHHFNQLSIKYHNNNLYTTLKNHNDINYNNHHIHPYKMLYYTNPNRTPPPKINPKPIEMYG
HRQRRPSW	14	14	6	21 68 39.5 17.53	0	17755_3	SYSSFGASSSSSAAVALTHHAPAAITLTEREECGVCRQSDRHRORRPSWHRHPFSTERQPTSLVT
TTPGSDSG	14	14	14	21 25 25.0 1.03	0	23583_3	AATEKRONVNYTTTPGSDSGSTAVS
TORQOPPT	14	14	3	37 70 57 13.57	4	34008_3	HSPHTTRHPRNEGTORQOPPTMOEEKASTTSPSSSLQPOGNGQTPRPPQHPAPANRPSITRQNSL
RPASAPCG	14	14	9	35 52 49 6.08	1	10880_5	PPSPSPSCSPVQRTRPASAPCGRRRPQSVTSAHPSRGAASVATPASSRPHRH
THRHRLQF	14	14	2	30 54 42.0 12.0	3	20835_3	STQQAHTPPAATHRGOGGEATHRHRLQFNPTANGIRPQRCQQTQRTPEQSPK

PPPSPPAH	14	14	3	51 55 54 1.7	4	22704_1	ADNMRRGQQAQPHSPHTTRRHPPPPSPAHFHNPREMAESPAATTQHTRRKRTL
PGVFETTG	14	14	7	40 44 40 1.4	0	2254_4	VACTPRVTDGPRLVDDRSTKVPGERCAPGVFETTLGLRLDDGV
YYTNPNRT	14	14	1	58 58 58 0.0	0	25282_2	HHTISYCRHLAINTQQQLNKLILTRHNTLYYTNPNRTPPRPKFKSGKMYGGDA
TRAPSRLR	14	14	5	26 39 34 4.17	1	20766_3	TSSATTTTTTTKAPITTTTEAPTTTTTTTRAPSLREID
PTQRSHTR	14	13	2	47 74 60.5 13.5	0	28075_4	LGAHDTISNPPAQPPHHQHVVHQSSQSSHTRSKQHNNHAQANPHSPMTRHQNFSHPVLSGTRSPTORSHTR
CSRVRSP	13	13	6	28 62 42.5 10.87	0	47271_4	KNGPKSRQHPIHAWRAPCSRVSRPAAECKERPTRIHSNGTVIPTADRHAASTRPLRHGTQANVV
LGSRGHOA	13	13	4	33 57 41.5 8.73	2	44466_3	QQHRSTADPKGOERGAAMPRTGITLGSRGHOAVSHGFLAAWRSHGGHDCAVPVCGA
TPTPIEPH	13	13	3	18 24 18 2.83	8	34213_1	ISIIYITPTPIEHLPENSAKIRK
VRHGAHPH	13	13	3	48 52 49 1.7	0	14609_2	TAGRGWRPTLVRHGAHPLPSKCTQQGAQSPSHIRNNGQTQTVVADYRCL
HAPYAAHH	13	13	3	45 51 46 2.62	1	40850_3	EGTPHEKRAAATHAPYAAHQAOTHSNRPQQAHTHALPQHQTDDGRGSRQ
KRWWIFGK	13	13	8	55 81 69.0 9.14	1	47226_7	GAFATGWSASSRQHENAGDAAKASAPLAESREKRWIIFGKNGSGDKNAGFSAGTAPSKVNSSSGGGNSGGMNGSVDDD
RWNQCDRC	13	13	9	22 37 29 4.37	1	11514_3	SLTLTLTRPPPOQPHKPRNRWNGDCDRTPTPSCKE
HLQFNPT	13	13	1	89 89 89 0.0	0	28618_3	HGRGEETQRHRLQFNPTANGIRPQRSQHTAHTEQAITQRATRRSSSPIHDAAPHSMRSORSAPHHQETLVRAPSLPLSHAHGK
RDGAQRIQ	13	13	5	13 85 41 23.09	0	31500_4	EWNERERDGAQRIQRESADNMRRQQAQPHSSHTTKRHPNEGTQRQQPPTMQEDNTSTTISPSPFPQPGNGKTPRPHNAHGAG
TNDRREKH	13	13	3	41 56 42 6.85	3	18200_5	RQEALETKMSREHAATPRRSTGGEDNRRKKQEAQAQHTETNDNRREKHRHSPTGRQ
SSHPPCRR	13	13	3	30 59 37 12.36	4	40395_4	ARKAHNASSHPPCRRRRHPPPSPSAHSRSPSEMAKTPRPHNAHGAGKPSIIHAAQNY
APHRHHS	13	13	6	13 49 28.0 11.78	0	38438_4	TSTGTTTTAYKSPNNSPGLPAAVTHTDEQQQKHSARNNTAPHRHRSIH
TTTTTTTT	13	5	1	56 56 56 0.0	0	27238_3	SDKRLSEEQANSESEDSTEETTTTTTTTTTTTQAPSTTTTTEAPAVST
HTPIHGS	13	13	5	31 50 45 6.43	0	16108_3	GEQSSQACTVRRAPRTHTHPIHGSRHTHTQPCHTNTQTMEEGEANRSQTQ
YYTNPNRT	13	13	1	50 50 50 0.0	0	41453_3	TLHINYQIFLPTYLTYTNPNRTPPRDNPOFWRNNVRGRCMNPFQI
TPSPKHLV	13	13	11	17 49 21 8.96	0	38989_2	ERINRKKGSKPSRMSPODRNRIKIPWTTAATPSPKHLVKPQDLRKG
PALKKDEK	13	3	3	69 80 69 5.19	0	39699_9	KKYKAISVPVKDEKVPSPALKKDEKAPTALKKDEKAPTALKKDEKAPTALKKDEKAPTALKKDD
TVHTPRDG	12	12	7	14 54 20 14.21	0	47468_2	TTRREGSRPSTVHTPRDGRHHHHPAHFHNPREVAESPAATTQHTRRKRTL
HRPROGQY	12	12	7	25 44 30 6.47	1	14796_4	DACGEVCSALRARAAHRPRCGQYTCRKNMATQREEYRASVTG
RGHHTHT	12	12	2	48 67 57.5 9.5	2	40936_4	PREMAKPRGHHTTTEQAPTRGRSGSRASSIHAATKFAIPAPSTGHGAQLNAVTRGASHPOERRR
HMSLRPHG	12	12	12	31 58 47.0 9.75	0	44838_4	PVTQRREEQKPVGDAAGLRNDEGRHSLRPHGRGPGRAEVGORSVYADAGWETV
RWAWHSHG	12	12	5	26 41 29 5.46	1	11506_4	VPQSQRWAWHSHGSLGVAPSRRTATPVGEAAEAHPLQQR
RLHQRRHT	12	12	10	48 77 64.5 9.16	0	7414_3	AHPGMPKSDVLSASNAITPIHTCRLHRRRHTNHKKHPPHVIQTFRHTATROTEVFVTPMHMLPTQNSPCTHSGSSS
PPHACSRH	12	12	5	38 60 46 8.45	0	42468_4	PTGPRRTPTRHGLSAPFVAAPHACSRRTQGRHERVGGCAGAVGREHPAPLAGRVQQ
DPHRRRMP	12	12	3	27 41 30 6.02	1	44229_4	KLSPDPHRRRMPWTAKECYVPGVLNAKEKMLDGVQLVDE
KECVPGVV	12	12	6	24 40 33.0 5.19	1	26822_4	SPDPHRRRMPWTAKECYVPGVRSSEKELVDGARVRDVE
RSRGDGV	12	12	9	21 32 29 4.07	1	41207_3	SAASHLLPSSFRWRSRGDGCVCVGPSPQOH
THRHLQF	12	12	1	79 79 79 0.0	0	37728_3	THRHLQFNPTANEIQPQRSQQTQRTSRQSPKEPHAEIHPHTTRRPTPSGRDTPARCLSPFFASCPREKAAPT
AVRDGRWS	12	12	7	26 50 36 8.53	0	1666_4	LRAVDGAGRADCCAERVAERCAVRDGRWSVARCDGGERERAGGAV
LSPRGKNQ	12	12	7	28 69 39 11.97	0	27355_6	CKTLNSFHVNKQCGTMEPLHNNVWRFRPERPINTARGHSEKATPPFVLSPRGKNQSTIQANNAHHI
QGTDSWRR	12	12	11	27 45 37 5.51	0	25928_3	DDHYGCTROEHNHLQGTDSWRRERPRPFVSGSAAAAACQCRG
PPPPPPPT	12	12	5	36 77 40 15.44	0	13382_7	TRQFIADNFAPVPCNEFLPAPTCSDAEQCTQGVIEEKKRQIQRREWQLDGRYHPSPPPPPPTTTTTSHDEE
SSTHHHCR	12	12	7	30 81 49 16.85	0	8940_6	PSPTQKISAHLKREGTKRNEGTPPSRRGPTLSSTHHCRAGRNSHIAESAPQCEPHLNVNNGRDAARHPAPFSSHSS
TAAGGGRS	12	12	11	40 75 55 12.84	0	11913_6	ESESTASRVQSASNSAVRSTAARKGRSSAEVRQPTQKKGAKNEGTSPPVKRPSRTPSVTPPVSAASAPARH
HHLPLQLP	12	12	4	74 78 77.0 1.5	0	26470_4	KEYKKRERDVGORIQRESQRHEERAAGPATQFTHETTSQGRHTTPAATHHAGEEDIIHHLPLQIPAAPGKWQKHGRG
GRGTEHSG	12	12	5	17 26 18 3.66	3	41465_1	VGQKKNGTGGRGTEHSGYSERAOTT
YYTNPNRT	12	12	1	46 46 46 0.0	0	19106_5	QSCNTIKELPHRYIYYTNPNRTPPRDNPLKQNNVGRCMKFPK
TNADNVRS	12	7	4	55 88 73.0 11.78	0	11995_9	TPARSTANPTPQFNHSTNADNVRSPLKQSTNADNVRSPLKRHSANADARTGSSSKSPKYSKKDSIITSPVDRAKLASNVP
RLAEAEKE	12	4	4	56 85 71.5 11.25	0	36997_3	VGSSMQQQSNDILWEAKORRMAAEKERKRLAEAEKERRLEAEKERRLEAEKERRLEAEKERRLEAEKE
PTQRSHTR	11	11	2	19 64 41.5 22.5	1	35871_3	HHAGGEDIIHHFPLQIPAAITQRTSRKQTLHPSRGTKISCHPAINGTRSPTORSHTRCITPAG
TSSRADGN	11	11	11	41 71 57 8.37	0	3601_4	FATGKRGSVMREESNLSSGHPORAAVTSRADGNQORHRSHPSPFQNTFPVQNSNNAVAANTAV
YITPTPIE	11	11	2	36 47 41.5 5.5	1	25378_2	NNYKHNINIHPNLMRLQPHLHLCLYITPTPIEHLVPKANQTK
RPPQARG	11	11	10	12 17 12.5 1.69	1	33730_1	TATTPWLSRPPQARGV
GTHRHKPP	11	11	5	31 63 43 10.46	2	776_4	QERKPSGNSAPSIIPASNATSKHPPPMQGMTHRHKPKPTATCTAANNTAASRWATSPRCLH
QQHSSHTT	11	11	11	12 76 30 17.73	0	37096_4	SRVAEFSLICGTEKEWNRERNRAQRIQREGVDNTRKGQQAQSHSTTRRHPPPPSPPAHFHNPREMAESPA
EAWDDQRT	11	11	4	25 50 36.0 8.98	4	28423_4	RTKQKAEAWDDQRTSHSYSRKAKHTHTGREERERERSPTQRNKNKETVT
TKQKQLRC	11	11	6	31 68 52.0 12.08	0	11107_4	NEKKQKAEAWDQRTHTLTPKQNTHTTEGEREKSRLNAMETQKQKRLCSAHAARQDGSAGCRAKE
AAARHAPY	11	11	3	35 47 41 4.9	2	29716_3	HEERAAAARHAPYAAHHAHTLQSTATRTHIHTPCRNANTQMEEGA
GTPTQDRS	11	11	11	14 41 32 8.66	0	9639_1	VAAANLPSTAPPRRTHLPQWWRPRTAGTPTQDRSCWTPPR
THRHLQF	11	11	1	57 57 57 0.0	0	17827_3	NGRGREETHRHLQFSPTANGIRPQRSQQTQRMQAITQRATRGSSSSIHDAAPH
HLQFNPT	11	11	1	71 71 71 0.0	0	45637_4	TTNSHRHLQFNPTANGIQPQPAAGTQRTSRQQSPKEPHAEIHPSSHDAAPHSTHSGERTVSSTELUR
QRQLEELR	11	9	6	37 62 52.0 8.79	2	30636_5	LRAENEQLRVENEELRGEHEHKTRGLQVEQAEQLDRLQLEELRAENEQLRVENEELRAEGE
ADYVGAML	11	11	11	28 72 36 16.51	0	4991_5	AESLYQEGFISYPRTEDSFSFTDDELREIAGVQRDNEPVDYVGMALDDSSHKFRPLKGGHDKAHPPYI
PWQGLKAH	11	11	11	24 38 31 3.84	0	40200_3	GSLPWQGLKAHTKOEKYSRISERKQTTPTVETLKGFFA
TATDKLQH	11	11	9	32 57 40 6.87	0	3418_3	VQRIFATEHSRVHRDTADKLQHSVAHHHRPSTSHHREHGHAEHNNQHPHEPQTRQH
LHGWRNWT	11	11	11	29 43 34 4.57	0	24798_3	QGRSRHGITSRPHSLGHWRNWTAPSPHPTQKAVSGWTSVAG
AQPHSSHT	11	11	3	21 81 26 31.89	0	43806_4	ERDGVQRTQERADNMRRGQEAQPHSSHTTRRHHTKTHNASSHPPCORTTHPPSPSSSPQPGNGKTPRPHNTHGAGNHTSHTRQW
GRHPQRHT	11	11	7	17 19 18 0.83	1	11848_3	AARAGRHPQRHTSHITLPP
PRPHNAH	11	11	2	42 68 55.0 13.0	3	16932_4	HNASSHPPCRRRHPPSPPAHFCSPREMVKTPRPHNAHGASEPSSIIHAAKPLAIPPSMGHGAQLS
TVQPAPSE	11	11	10	49 51 51.0 0.92	0	25717_4	PQTVPQAPETKAPQSPCCDRKAGNARGLAPHFRGRPRDKDIAETRT
PCGEASHS	11	11	8	36 47 40.5 3.9	0	23251_4	LFPHPVELQRNNHSHSGEVPSONPCGEASHSRPTAASRCRWLQST
SGVRORGC	11	11	3	48 57 56 4.03	0	41401_3	TSDRGRRQAIVCEWHAQPAYTSSAHSGVRQRCGSGWHCCSTGTAAAPKETNVRSS
GRPRGRQL	11	11	2	45 80 62.5 17.5	0	34762_5	EETNDRREKHGRPRGRQLHTSPRTHAPSQQCSAGEKIRPSIQTHTHMEQAITQRATRGSSSSPHDAQSMRSHGV
RAPEQVQK	11	11	6	26 75 42.0 15.95	0	28260_6	PLNSTEMAIKDRKPVSTRAPEQVQKPAKPVAPAAPAAPVAPAAPVAPAGNEGTAAREKGYVGTNG
SSDIEPFF	11	11	4	36 58 47.5 8.17	3	15254_4	VISKQSSSDIEPFTSADVMAEEDSPONGNTDDPAPQGSNDVLESVHDEPSNASTL
MHRTPHTT	11	11	2	31 55 43.0 12.0	0	28614_2	NHMRKQQPGMHRTPHTTHTDSNPRQQAHTHRDGRGRSKQERHLRHRMHIRE
NPSRAHA	11	11	8	40 63 50.0 7.22	0	38829_5	VRQLKPEAHRSSVRGDWEPTHTDSTLNGKRPWHWSASAGFRNTDGDYDHCNCSPSRAHVGA
RCSQRHSC	11	11	8	38 49 39.5 4.79	1	6612_4	SYKPVSETDPRPSRRCSORHSCSVSSAPTPCCIVARSRPSHQAQSGT
DQGSRCFS	11	11	10	50 58 50.0 2.4	0	1254_4	RDMMNEAQLTRGQKNIGVRLLFGDDGQSRCSFSQIAETVDALNEKWTAEFRQIDTEH
QQQQQQQQ	11	3	1	63 63 63 0.0	0	9422_3	LPPPPPPQQQQQQQQQQGTHSHKKARYEREEGGOTROYLHYQRHQQGQTOEREMRRGGGR
RNGKRDG	11	11	8	32 52 33.0 7.34	0	15179_3	VPMPDFRNGKRDGTGGEKGWQRQQPQRQSTFPVSPRGMVHRGNLDPTSFKN

RRPDAADG	11	11	6	26 70 56.0 14.72	0	45371_5	GEQHAVVERGGVPDARRDDDDGGWRRPDAERHPWRRRRPDAADGGAAAVSVGTRPATRHAPELRAGVDG
GRGRGRGE	11	6	6	34 60 50.5 9.18	0	29039_4	FLNDEGEVYLKRYLFLPHDAVPNTHKAEYKVLEREGRGRGRGEGRGRGRGEGRGRGRGE
LYYTNPNR	11	11	1	43 43 43 0.0	0	14807_3	LSTPLFLTGLYYTNPNRTPPPRKIRDFEPNNVRGRCMKSGPKI
PRRQGGRL	11	11	9	23 39 33 5.29	0	28284_4	QPRROGGRLLHHWGADRHDDCCDSOGLHLLGACAGOGGA
RGHHTHTT	10	10	2	36 55 45.5 9.5	4	43000_3	ETSTTISPSSFSAAPGWKPGRGHHTTTEQANPHPSITRHKITCHPAINGTRSP
VRHGAHPH	10	10	2	37 49 43 0 6.0	0	31916_3	WRPRTLTVRHGAHHPFLSKCTOQAGAQSPSCHRNGENKQSSSRITGIASSR
RRYQPRVH	10	10	10	25 50 29.5 9.1	0	18483_3	AEGGPALPEQRDSCGAGGDGVLRCVASRRYQPRVHPORPDPAKHGPPR
TQRQQPPT	10	10	2	66 68 67.0 1.0	5	16583_3	GQQAQPHSTHTRRHPRKDGTRORQPTMQEEKISTTISSQLIPTANGWKPKPAATTQRTRESEHPHQ
SHGKVVMT	10	10	5	37 47 44 3.97	0	16001_3	GEAAEVHHVHAGARWTHAEGSLHAVGSSDRLSHGKVVMTTKVTS
RLMVLTS	10	10	8	25 64 45.0 14.33	0	25453_5	EGMLKVEKGKPPQSWTYKAVGDLEKDDGVSQSGAPRRLMVLTSKDGWVPYVWKWKENKSTRD
PCSRATAT	10	10	3	26 45 26 8.52	0	38620_4	SQRRVAWVSHSGAVGAPCSRATPVGEATEVHRPQQRNVPQTITP
THRHRLOF	10	10	1	56 56 56 0.0	0	4949_3	HTPPAATHNGRREETHRHRLOFNPTANGIRSQPAAGTQRTPSKQSPKEPHAEADPH
ANGIRPOP	10	10	4	22 73 58.0 20.61	0	11440_3	QSAGERIRPSTQOQYRTRNKQSSSTNGQEQEAHSHHRLQFHPTANGIRPQPAAGTQRTPSKQSPKEPHAEAIPIH
MEEGEANR	10	10	2	41 53 47.0 6.0	1	11391_3	NNTQMEEGEANRSGTQGSTACTYRESGSPCRHPAWRVTRHTRGGAAHTATQKH
NGTRRSAP	10	10	10	38 67 52.5 8.61	0	1195_1	IQDATRRDIVHTAGNTRRSAPHGSGGAGGLLPRAGTKPGRPLPTGHHKVPVPRACROCTPGLFHSL
LVHGKCSQ	10	10	8	27 79 40.0 15.02	0	9051_2	EEGEANRSGTQGSTACTYRESGSPCRHPSRRIHTHRGGAAHTATQKHSNNEKYPRTPCGLLVHGKCSQYSPQALHLQ
RGQQAQPH	10	10	4	34 55 37.0 8.58	2	27640_4	EKNGTRGRGTGHSYSGYNADNMRRGQQAQPHSPHTTRHRPRKEGTORQOQPPTMQ
SSSSTRAQ	10	10	7	45 96 81 18.67	0	29592_5	YNSHTEQHRQONTSSTPHPHARDSSGCRHSEHTTHPPAQPHPHQTHVQHSQPPSRTRTSSSSSTRAQGITRRTPHGAAPSPSSSPVVELSN
ARRKEFHQ	10	10	8	35 89 61.5 16.84	0	18133_5	LRLQLEEAARRKEFHQTRGEOEPRHGRHERONNGTRNDQYGGNNSNRDRVEWNNDRDGRDTRMDTRLURNEGVSRRNRPRAERNN
GGFGQOAG	10	6	6	42 60 51.5 5.71	0	45025_3	GGFGQTAATGGFGQOAGGFGQOAGGFGQOATGGFGQAAGRGTAAGFGQONTGTTGGFGQOT
VKGRESVS	10	10	10	37 96 67.5 18.42	0	8018_6	QHTPTNFADATSTTTTTTKSSSANKVKGRESVSQSVHRMPTKPTSIVNRNPRPRNDASQGERITLREHASSLVKSDRSRDIEGHAAEGLMAS
AHRTLRAW	10	10	9	14 24 15 3.56	0	27125_3	ASTCGCLERPAHRTLRAWHGEGRA
LYYTNPNR	10	10	1	40 40 40 0.0	0	15614_4	KQRINLTYPMLYYTNPNRTPPPRDNPLKSNQNVNVRGRCMN
PAVGGFGS	10	6	6	20 45 27.5 8.55	0	5824_3	HTSTPAVGGFGSATTTSTPAVGGFGSAAHTSTPAVGGFGSAAHTS
QGRRHERV	10	10	4	33 38 33.5 2.06	1	24699_2	AQGRRHERVGGVRVEQLGGKHHTAPLTGRVOKKQLKMEE
ARRAQPAL	10	10	8	23 62 39.0 10.99	0	43321_4	EHSPQRSRLAGDCTSVDGRDGTGARRAQPALSHRQRCLTQSCRSERESGDDGAARD
PFGQAAAG	10	3	3	50 76 55 11.26	0	36781_8	DKPPFPQAAAGDKPSFGQAAAGDKPSFGQAAAGDKPPLFGOGTVFDASRSTVFANAPGVAQ
PAANTQRT	10	10	3	30 70 60 17.0	1	18696_3	KSHGGETQRAARETSAHRQOPHAMDEEKRPATSRTFSRPTGYGHIIPAANTQRTRESEHPRPQPHVEVDPH
PTASTHRE	10	10	7	28 61 34 10.5	0	31081_5	DGASVTARTLHTPGATPPALTAAPWDAPFTNKSHETRHSQRLPTASTHREVRQAAREG
DLCGGRRG	10	10	9	28 43 37 4.05	1	10399_4	GRDCGGGERFRFDVGGVLRWCDCGGRRGGRGDVDGEAORD
EVAKPEEK	10	6	4	30 73 48.0 15.88	0	19267_3	SRVSRENAIDGAKQKENTKPEETRLVAKPEEKKEGAKPEETKLEVAKPEEKMEDTPEAKTSEEPKPKET
PTQRSHTR	9	9	2	47 60 53 5 6.5	1	24666_3	PQLIPTANEKWQKHGHHTHTERAPTTPATRGNKISSHPALNGTSPQRSHTRCTTPA
YITPTPIE	9	9	2	22 29 25.5 3.5	2	38649_2	TLMYLOPCYIPTPIEPLHPDKITKSAAK
TGKSRPST	9	9	5	37 73 55 13.26	0	37761_5	ARPRTGKSRPSTQKTYTPEVTNPSFOIRNKNHTASHVQGRPLSHPRGDKVCCPALSGMRRPTPNGHTRYP
PQQSRSPF	9	9	4	30 40 37.0 3.94	1	9951_3	SKWDDKKHQFGNRPFPQQSRSPFLTAKACCGAADDGTH
NDSVRAGD	9	9	9	56 88 68 8.97	0	24806_5	NDSVRAGDRPRGGVGTGRDGAASKHGRTPGRQPRHLRGIPDAARTAAEHRQGVGRDGVVPAGHVAEVDQSRKPAVAAQARV
SQVSRGWN	9	9	4	39 58 48.0 7.6	0	27910_4	DTCEGWRCRSSQVSRGWNATFQFIGDNRKQHGLRRTPTMQLQDGPICASSSPGEAE
ARHAPYAA	9	9	2	38 48 43.0 5.0	2	21239_4	EQQGRREVQRTHTEGTPHEERAARHAPYAAHHARHTQPCRNNNTQME
AQRNGRPL	9	9	6	39 59 54.0 8.04	0	14571_4	HWRPAHRTSSSLAPRSRTHAQRNGRPLAPOPPARGSRAGSDTLHRGTGQSMTHSTS
QRAKDGOR	9	9	2	38 69 53.5 15.5	1	19417_5	PTTASRAGHHQKFQAYQRAKDQORTSASSNDSRSPATLLQSPPAEPRSTGGPAAHVGMPSSAT
TPHSHGRA	9	9	9	25 64 33 10.96	0	361_4	IPKKGORLPDLRSSEDLMLASEPVAKSSTESVDSTPHSGRAWGAKEARPPRPREKOPTPA
THTEETND	9	9	2	39 81 60.0 21.0	0	45283_2	THTEETNDRREKHRHPQPGPQRLHTSPQHTASSHPGTRKRRESPPPPYQSHGQONTATAPQPTHKAQQAITQORATRR
SPREMAKT	9	9	3	43 64 56 8.65	1	33371_4	HRYSPREMAKTPRPPYQPHAPANPLMLTRHKIPCHPALNGTSPRTOQRGHTRCTTPAGKETRGA
THRHRLOF	9	9	1	54 54 54 0.0	0	23440_3	EDSALHTKAHTASSHTNARGEATHRHRLOFNPTDNGIRPOPCSQHKAAQAITOR
RYFGVCDG	9	9	9	48 71 69 8.14	0	29299_4	LREARRAVERRYRYFGVCDGSDAACGGGGRGKSDATKSTCWRKGKERSANCLDAPVGPATMNTADTQETG
NTQMEEGE	9	9	5	24 54 28 10.84	0	19219_4	RRGQSGMHRKPHHTHTHSNRPQQAHTHTQPCRNNTQMEEGEENRSQTGST
NTKQOHRE	9	9	8	56 90 69.5 9.83	0	26623_6	EETHVAVEGAAARSKVPVFSKAASFOALNTKQOHREGRQGRQGGKDRGDKTEGKSMPTAKPAENSGATISRPVRAPAVPSSQHPEK
QPRGRQOL	9	9	2	27 51 39.0 12.0	1	41942_1	RRKGEKQEAQHTTEETNDRREKHRHGRPRGQRLHTSPHTRTIAVLVSRG
GTASANPN	9	9	9	12 62 21 12.64	0	18971_3	YORDEKGRAAPRRQSVRHLPADQASCAGVAHRDAYVLVAPECGTASANPNTAVNPLPQS
HTLIAPVK	9	9	3	40 51 43 4.64	2	9986_4	QPQNKQGNKITEAWDYQRTHTLIAPVQNTHRGREREVAHSMQKQNRN
MHRTPHTT	9	9	2	25 52 38.5 13.5	0	13572_4	GQQPRMHRTPHTTHTSHSNPROQAHTHTALPQQQHRWKSEKQTGAHKAPP
RMGTARAG	9	9	9	27 43 38 4.32	0	17242_3	FQQAASLVPODTQKEHRRNRMTARAGDRDRDTRASPCIRSA
SPPHSART	9	9	6	46 58 53.5 3.73	0	8592_3	LHPPTQKGVGMKGSPPHSARTKEYAAATPSTSHCKKEPCHTKRCPQREGSGASKST
KPVAKTAA	9	9	9	16 42 35 7.6	0	16817_5	SAAKPAKPAKTAAPVAKTAAPAKKPTVPKPAKPAKAAK
LYYTNPNR	9	9	1	37 37 37 0.0	0	24058_2	SPHTTLNHHHLYYTNPNRTPPPRKIQKFHKIMYGGDA
TRAPSRLR	9	9	2	34 45 39.5 5.5	0	8166_4	YEVAAPENSIEEKSNTMPTMRDAATEASTTTTTRAPSRLREID
PRPPHNAH	9	9	2	43 59 51.0 8.0	4	21081_3	TASTORKOROQPHTMQEEKSTTISPSSFPQPTNGQKPRPPHNAHGASTHTHTSHTRQW
IDASPPE	9	9	6	21 28 25.5 2.54	0	41728_3	TALRQIDASPPESTATAAPCVVLPQAQSL
LGSRGHOA	9	9	2	60 62 61.0 1.0	0	38742_3	SERSGGSRAQTRSQHRGTADPKQGERGAAPRGITGLSGRHQAVSHGPPAAWRSHGGGN
FFQHHDAA	9	9	9	22 52 35 9.27	0	21133_3	PFLDELKREYSHDTDTRVAAPYFFQHDAAGCKTSRCRHKEYEKENMMORME
DVGPRHVD	9	9	6	21 52 32.5 9.19	0	48407_5	RHVDPDHFRSTTODAYRPVDPSPYKRALPQEEQEDVGRVHPDOPHFRSTTOD
RRCNPMSM	9	9	8	34 82 60.5 17.88	0	34748_5	MCSQRSYRKSQQAARWEYDHCTYPSLASROWWRNRCMIARRCPNMSMSPPLAPVSRSDAQQODFMLHPSSCSRTVKSTHYPSSTAS
VCGAAGVC	9	9	5	15 41 15 10.23	0	11137_5	PKGAPAKKKHSEKVKESQNYFLKAAGCGCVCGAAGVCQP
RPSRTVHT	8	8	7	14 62 17 22.83	0	45130_3	RDGAPRIQREKEQTKRGEGRSPRSTVHTPRGSIHAQTTPAATHHAGGEDIHHLPLQIPTA
TKORRRNQ	8	8	5	65 79 78 6.71	1	41658_4	LHGQKGIKTAPVTKHSGTKGQPHAINPHRSTADTKQRRRNQVSDTIALSLPTQHTVTVPSGGIPAKSHMTGTAPV
GRHGGSPP	8	8	5	31 52 40 7.83	0	42035_4	VLLKWRQEQQWQYIRHVSEEQSRAPTQORAGRETIRKRGHGPSPTKSSNTA
ATPRRNTG	8	8	5	16 16 16 0.0	1	47590_2	ATPRRNTGEDDKKKK
LCRTGRQA	8	8	8	24 44 34.5 6.3	0	18605_3	QRRVGACDCQHLRCRGARLCRTGROAERTNKTGRQQTLSFNVS
TPTDEEKG	8	8	7	40 87 56 15.65	0	16838_6	TDEGHTRPRTPTDEEKDWRRISRGGQFLAVGRGSTANWSKPLQVQDNVGNVGVAFGSLGRAEGLDDGDFTEVLRHPTIPTFVD
HRQRPPSW	8	8	2	22 51 26.0 11.52	1	2890_3	ARSNALPQSLPQRDKKESVDVCREHSDCHRRRPSWHHPRSVSTQORATPE
AGRPRASG	8	8	5	40 85 61 15.78	0	41505_4	KRERNDAKKIULLTDQVAGRPRASGOKKTSRSGRLATFSFSSQCSAPPTNHLFGPCSCGQTSRCWISRSRPHRSEGKKRD
PLUSTPG	8	8	6	36 68 45.5 11.88	0	15522_4	TPRDGTRHHHTPLISTTPGKWKAPRPHNTHGASEPSVGNDAATKFTLPSPPHWNTEPNPTQOEK
GORNASS	8	8	7	31 64 57 12.0	1	22915_4	TSRVGVCYRKQWYQEAKHGORNASSNDNRGSSATPLQPLSALAEHLRRSTEDSAAHAVGTPPP
PLRTPCGH	8	8	6	33 89 41.0 20.34	1	18484_4	KKERSTSSSSTSFSOPTDMATSQPAQASHSTTSHPKPLPNHPRRGDMNLHWHPALNGMPLTPCGHTRCTKPAEKESRAAHSLLP
LTVRHGAH	8	8	2	34 43 38.5 4.5	0	36268_4	LPPRRNVTPHRSRVPWHPLTVRHHGAHPLPSPERTQQAQGS
GVQSGSTT	8	8	5	21 47 37 9.17	0	47611_4	VYGSTANGNWYMPGKPITRVWGNHSRSHGHVGQSGSTTAAIEGKEV

APHRHRS	7	7	3	15 30 17 6.65	2	18695_1	THTKDQQOQKHGAGNMMAPHRRHRSRHHVS
WRITARWWS	7	7	7	20 29 27 3.68	0	20834_3	ISFDVAWRTARVWWSHSGSQWDRPACKSS
NNHARWQH	7	7	3	24 38 28 5.89	1	28266_3	NTRVDDTTGWQRYMRHGPDLNHARWQHQTQSKTKQKHGG
TRAPSRLR	7	7	2	34 35 34.5 0.5	1	36599_3	TSITATTTTAPTSTTTTTEMPNTATTRAPSRLRKIDG
HTHTQORER	7	7	2	63 66 64.5 1.5	0	11271_5	NSTPRHNSQATATHKQPKQKTRMPLPSPNKRPHHTQORERGREISVWDGKKKEEDELTRNAPL
PPHNKRTH	7	7	2	59 64 61.5 2.5	0	4237_4	SPRARGFGPPHNKRTHRQOSPNTNEREREETHRHRLQFSPTANGTRPOPRSQOTQRLSETSGS
PAANTQRT	7	7	2	55 59 57.0 2.0	1	14176_3	RGFGPHNKRTHRQOPPAAMEEEKRPNATASISSIPRTRHGHSPAANTQRTPSKLSPKE
LGSRGHQA	7	7	2	31 55 43.0 12.0	1	23526_4	RAQTOSQQYRSTADPSERERGAAMPQGITGLGRSGHQAALHGPPAAWRKHRRGHDG
VAPPSPHT	7	7	3	34 59 55 10.96	1	31850_3	VRAEGSSTHTARHGPRTRVAPSPSPTHFKNSVQPRPHARNPNHARHNNNASROPPTOQK
APSLRLREI	7	7	2	25 36 30.5 5.5	0	4225_4	TTTTTTTTTGAPTATTETPTTMTAPRSLREIDG
LGEEEVPL	7	7	2	45 52 48.5 3.5	0	4251_3	GKTTARSTSTWEPKHEQLALTLQGNKASVDVGGEPLGEEEVPLTERPPEVL
WRQRHHHR	7	7	5	16 34 30 6.37	0	26765_3	GSWRQRHHHRAPAQYQQCAYCPSAKDGVHPFLFP
DRRDYGDOR	7	3	3	43 81 55 15.86	0	41451_5	VEPYCRNCGRNHLSRDCRSGPRNRRNONDRREFVRDRYNDRRDYGDORRDYDORREFVDRDRDNDRRD
VTRWWKRP	7	7	7	27 50 34 6.87	0	46260_3	SDRATTSSTVTRWWKRPVLPNHRTPVPPSARSSPRGVTAESAHALYVSS
RNRDRREG	7	7	4	26 42 29.0 6.22	2	21573_4	ALSASRLSQRPHTTTEHNRNRDRREGIDGVSGHOLSSSSLLVK
THRHRLQL	7	7	4	29 92 44.5 24.42	0	9732_6	CTSRKTCPTMTHNGKKPLKRRALSQRQCSAGSERIWSSROOTNTPVITHRRGGTTHRHRLQLNPTNGIQPOPLQRTTSSKQSPKEPHAK
TGAGENEA	6	6	6	58 74 63.0 5.26	0	14508_6	NVSSLLSGELTGAGENEARSSGKGGNISASERTISGNLSRVOSLSNGSVSRVGNRESVSGRLQEDGGASLIS
LAKQOREE	6	6	6	33 66 47.0 13.44	0	19284_5	RRAELAKQOREESRARKEELQRQAEERKKKEELQAEERLLAEARSAEGEKKALAEKVRTGKE
YITPTPIE	6	6	1	42 42 42 0.0	1	15927_3	MPHIILSYNHKKFKTNINYESIYTPPIEPHLPENPAVSLK
TLKKGGGG	6	6	3	43 107 90 27.07	0	37195_3	TLKKGGGGGGKGTSGARQKQPKQPKQKQKHQYHHHQHQHQROQHQLOQGGMPMLMAHQVLLPSSSPTEGVFFPARGSGSGVLPTVPVVFQQQOQPPHLPFPQ
AEGTHQOH	6	6	1	42 42 42 0.0	0	3352_1	NFRSIPAGTAPQHAEGTHQOHTASHHHRRHSHCHLRAASNT
KRRHDPTV	6	6	1	96 96 96 0.0	0	48227_4	TPSQGRNKEKRRHDPTVPPRANLLHPPLPHRKTQSHRRVSTMCTTPPCEWRARHSTQSPRSIFPFLKCCVPAQTQKHNQSIHKPNQHPSL
GTHRHKPP	6	6	2	33 94 63.5 30.5	0	5044_5	AGRDHAKNPQKHQKQTLQRRSCDSNAKSRAPTOQPPTHTKRGKGGGPKAPKINRGMKMKQCCTLHTYSIKCHRSSKHPPTQGTGRHKKPPN
RGRGRFGD	6	2	2	50 56 53.0 3.0	0	6845_3	VENREENGYNFGRGRGFGDGRGGRGFGDGRGHGFGDGRGGRGFGDGRGGRGFGDGRG
KASSWMHS	6	6	3	31 35 33 1.63	0	25340_1	DRWQKASSWMHSRSGSFRSDCWQTHASSRTCWSPC
KECVPGVV	6	6	3	25 34 29 3.68	2	4022_3	LHYSGHKGNETLSPDPHRRRMPWATAAKECVPGVV
EARRHDGS	6	6	6	15 19 15.0 1.89	0	28959_1	WGLREARRHDGSRVSPSGE
EKKLSLGE	6	6	5	36 88 72 18.62	0	45297_5	EKMFIQMIPVTIQTILWKEKRAQAEEDAATKGAAEEEPKSDRAKITKRDVTPPEKLSLGEKLLKQAKERKREOROERDGTATNEEG
WANDDRHC	6	6	6	25 72 45.0 15.85	0	19430_3	SHRRSOPRSLAASGMSPSRDGHQVASKNTFRGTHWANDDRHCMMCPGSAALHKSHFASSSSSGINTPVH
RDGAQRIQ	6	6	3	36 58 40 9.57	0	46128_3	SQIRGTETKKEYKNRERDGAQRIQESRQHGAEAGPAAHFTHHETAFTQGRHTTPAA
KWQHPPSSR	6	6	4	47 57 50.5 3.77	0	13489_4	SAQTMNTSTQGSKRRDTAATTAHNHKWQHPPSSRSPHTMRQLESVPEKPLQDHSHTALQTR
LRRGQRSD	6	6	4	30 43 34.5 5.32	0	15763_4	ENAMPLHNTAGWRQSGSERFCTLRRGQRSDKKACSLGTPNV
HRLQFNPT	6	6	1	43 43 43 0.0	0	24387_3	NKRTHRQSSANARGEATHRHRLQFNPTDSGIQPOPAANTQRT
LTVRHGAH	6	6	2	46 42 39.0 3.0	0	43886_3	TSRLTVRGAHAHPNPKCNOQVAQSPSCCHNGEQKQSSRT
RHPRNEGT	6	6	1	83 83 83 0.0	0	9415_4	RRGQQAQPHSSHTTKRHPRNEGTORQOPPTMQEKTSTTISPSSLOPQGNGRNPRPPQNAHRACEPSSIIHAVKTSCHPAIN
AGWREERSG	6	6	3	24 64 43 16.34	0	18729_4	GCAIAGWREERSGGKCSAAVKGAGCLAVEGISRPGQHRSCRAPSLSTNPNPNPNPNPNPNPN
HCRTGDRS	6	6	2	27 54 40.5 13.5	1	46007_4	PTPKMSAHLHKREGTKRKGDMPSPRRGSTLSSTHRCRTGDRSHIAESAPOCA
ERAQRDAG	6	6	4	39 50 48.0 4.32	0	23321_5	CGGERFRDVGVLWRWRDLRGLRGGRGVDVGERAQRDAGELPVCGRSEP
HLHLQLIP	6	6	2	46 52 49.0 3.0	2	36130_2	THHETASTQRRHTTPAATHHAGGQHHHLPQLIPAATTQRTSRQTLHPS
IPAAPGKW	6	6	2	21 31 26.0 5.0	3	17951_3	IPAAPGKWQKHRRGHNAHGAESP

SLHPPTQK	6	6	4	34 48 42.5 5.58	0	4331_4	PHSLRHLHSLHPPTQKGGAMGKRSPPHSACIKEYAAATPYTSHCNKEP
DSHRTGTH	6	6	6	27 51 39.0 9.09	0	15028_4	THGGNRDVASAGHLNAKGVPHASIQWRDSDHRTGTHAASPSOHTKTRFPN
RTPVAPPS	6	6	3	47 70 56 9.46	0	36937_3	PLATVRTPVAPPSPHKHAKVGGEFAAATATRRPNNTHRNSGASROSPTQQNLPGSWPDREIYGNQTS
AAGPAAO	6	6	2	32 61 46.5 14.5	1	18406_4	GVSOLIGETERKKNGTRGRGTEHSYSERREROQHEERAAGPAAOFTHETASTQRRTHTPA
EALGTLR	6	6	4	23 46 41.5 9.03	4	42247_4	DGRRRVYESADKGESWTEALGTLRVSWGNQKRHEKVDGSGFSTAT
RGDERCGR	6	6	4	16 39 19.0 9.42	0	8955_3	MESNRGDERCGRHTRHGHTRGENSQAQCTVRTRPRTHT
KAVKETO	6	6	4	46 87 78.5 15.88	0	4085_6	KEIGAMPKGGGTNGKHQHQQQQQQGGKKGSKKSDDEFDALLAKAVKETOALPKAENGHHQKQNNAGKQQQKKSATEEL
PALNGMRR	6	6	4	30 41 37.5 4.39	1	41886_4	APPATRRGSSAIHDATPKCHCRPALNGMRRRTPRGHTECN
GIRPOPRS	6	6	2	30 79 54.5 24.5	0	37468_4	HTLPEINHGGQGEATHRHRLQNLNTDNGIRPOPRSQHTHTTKQAITQARTRGSDSSSIHOAVPHSMRSHGVHFFPNRT
KPAAHQVR	6	6	6	31 43 39.5 5.24	0	36345_3	EAFHRNAESRRKPAAHQVRHSPSSASRRPALPLHCAAFCKRTGA
HSTSKQOC	6	6	3	40 47 45 2.94	0	39568_3	TAQESLRTQPRRHSTSKQOCOEASIAAGSACFHKHANRPAGNLAPS
SIGTGRGT	6	6	3	35 55 49 8.38	0	1024_3	GAGRFHWHMLNKTSRASVPDPESIGTGRGTTPPKDVTDAAPQKHPSPEEKEVASGA
HRYSPREM	6	6	2	58 68 63.0 5.0	1	45727_4	HRYSPREMAQKPRPPOQHAPANPRPSMTRHKIPAPPSMEHGAQLNAVTRGAPHPOENGVLRLSLSS
AGWQRYMR	6	6	2	22 45 33.5 11.5	0	26823_3	VWVSVEAEKRSTGRHINTARGTAGWQRYMRYPDINNHRWRHT
PPHACSR	6	6	1	59 59 59 0.0	0	22701_2	QKQSKTRGKKCVLRPTGPRRTPTRHGSLAPFVAAPPHACSRGAQGRRRHVRGGVRVE
VFGAPST	6	4	4	32 49 41.0 6.18	0	30000_3	PSSTAAPPAESPFKNVFGAPSTAAKPAESPFKNVFGAPSTDAKPP
PHRSTRVG	6	6	5	31 54 42 8.9	0	19242_5	NFPHRSTRVGAPRADCSSSSSCKERRQEWQDSRASESQASPMLETSRPLL
TRPCGOY	5	5	5	30 46 31 5.98	0	28990_4	AAHRPRCGOYTCRRENMATOREEYRASVTGSHYTKTAAFLERTPAV
TRSPTQRS	5	5	1	68 68 68 0.0	0	31703_4	RHHLLQLPIAAPGKWQNPAAATTQTRSKPHPEAEADPHSMTRHKISSHPALNGTRSPQRSHTR
TROPVTPH	5	5	1	53 53 53 0.0	0	37790_4	GSNSVQGRKCTROPVTPHRAANPTVLTEAEVAAFRSRTHCRHRCHSPAT
PORQLHTS	5	5	1	69 69 69 0.0	0	13498_4	RAGTLEKVRROKTTSSTTTTEETNDRREKHRCPRRQRLHTSPRTHAPSQCCOSAGEKIRPSQHTT
IRYFPTST	5	5	5	36 83 72 16.04	0	4555_7	TMSSTLCTSCSTPDDTLQAIYFPTSTSRMDADSADDIKGRSSRKNKHDEERKRAQGDGHVKLYLESFYDAEDDVEDDGS
YITPTPIE	5	5	1	28 28 28 0.0	0	39925_2	LTHNQTTSYTPTPIEPLPENLAFWPK
PSPTHAS	5	5	4	53 61 55.5 3.42	0	17904_4	AAKDKRKHGRDNQAGAGRTTKTLPLAGATAPSPTHASDNAAFRPTKGRSPSTQKTYTVPE
QRSRLAGD	5	5	3	39 50 46 4.55	0	32878_2	QAAVGNIAREHSPQRSRLAGDCSTSVGRNGTGAGARRAQPALAHRQRCRT
RFGRGRTT	5	5	5	40 51 46 3.72	0	3356_4	PAESARFGRGRTTACFSEGCSSRESSEASMWRLPSTLKEKSRSTSKFTAL
AEGTHQOH	5	5	1	41 41 41 0.0	0	15398_3	STTAGTAPQHAEGTHQOHTASHRRHHRGCTQSREGKREDE
KRRHDPTV	5	5	1	86 86 86 0.0	0	25435_4	RKEKRRHDPTVTPRANALLCPPLPRLRRRPSNRRVSTMTCTPSCSERWARRGQASPRSLPPLKYCVPAQSSKQNGRAHKPNQHP
RQSHRVRP	5	5	4	39 29 29 54.6	0	4162_3	RQSHRVRPPRSKEGYFCPPGRQSSTAEDCWFRAFQTKH
RSRHPAFQ	5	5	2	25 56 40.5 15.5	0	8050_5	VDRRAGRPAGSHRSNHEFGEVPTYPEADAHDCVRSRHPAFQTARPSLRHRCSPDQH
PAQSSKH	5	5	4	28 36 31.0 3.2	0	34013_3	SPRSIFPLKCCMPAQSSKHNSARKPNQHPHSLKH
ISEVERNI	5	5	5	51 68 60 6.02	0	42910_4	ERNEEVNRRISEVERNIAEORLHSHNQAEFNKALAEQKRREAIRDKEEDTRGLEIERYHLEGDFLN
QSQRGSG	5	5	4	28 47 35.5 7.02	1	28239_3	QSQRGSGAAQPSFVDTCDWRRCRRSOVSRGWNTTPIPOFQDGNRKG
RSHGKWR	5	5	3	16 44 31 11.44	1	25059_3	VAEVHREHDGAGKWAHAEGGGLHAGVASSDRSSHGKVRTTRAT
KFTYVRNW	5	5	5	24 31 26 2.65	0	45287_3	SLRKFTYVRNWFTSQGQAGNLSGADYDGP
PAHPHAAQ	5	5	2	37 73 55.0 18.0	0	40683_3	TRIPAPTRRLRHPPAHPHAAQSGSQSQRLQSLAVAGAPIAHISFLAWHRDRCDSLLSPSSPCCSMPSTSSQ
AARHRPAP	5	5	3	28 53 30 11.34	0	33222_4	TGDRSQAESAPQCAPRHRVNDGRDAARHRPAPFSHHSSIACPKHVANNRVR
LDLARDYK	5	5	5	37 47 42 4.26	0	15031_5	NINSTRHALDALDYKMRCFIPSTIAAFGDKCGKVNTKDDTLNLPST
VVEROREH	5	4	3	41 61 50 8.18	0	30333_3	LEAALLEDEAPHAVVEROREHRLGLDHEAEOPLAARKKRKGKLPQDGLFAFDDVEPDRE
HLPHHHPF	5	5	4	31 52 42.5 9.57	0	47502_3	AFPSRTRNQHHLLPHHHPLVLPSLHSQQSSVPLAHVGYWRFSPQLSSPVL
AERRERLR	5	5	3	32 59 51 11.32	0	22339_4	MLQRHGSQAKALAEERRERLRVDRYGTMRDASQDTTERNSSGTGRPSTQPALNDRWPRP
GSHSRCRL	5	5	5	36 60 51 8.13	0	37244_5	GNAGAGRRWCIGKKDPSLEGTPRAGRADGSHSRCRLHGSDDGGVRGYGLVPRELEHGGA
LRVLSLPH	5	5	5	38 41 40 1.17	0	1978_3	LRVLSLPHIPSHPHLPREKNKTKQNPDEKKKGGRGRGN
DEIPTSTW	5	5	5	50 75 61 8.21	0	10698_6	SMPVTPPLSEDSGRNSRADEIPTSTWRARRWSDFPYTROSEKNPKVNNQPHVFPQQRNGTMGAVNCEEFDAPN
GEGAERAC	5	5	5	14 19 18 1.94	0	38528_2	GGHCPCGCGGEARACGGC
HRLQFNPT	5	5	1	41 41 41 0.0	0	9621_3	GKATHRHLQFNPTDNGIRPOLTANKHNTHGASNPKSHQ
IPSAESA	5	5	3	42 62 43 9.2	0	2405_4	HSHOQRQPROTAEAPTDDGHQPORHARHPSPDARIPQROTHSSCLLSFCRPPRPIPSAAESAV
AAEAAKAV	5	4	3	34 62 41 11.9	1	40592_4	MKVAEAEKRAEAAKAVETEKORAAEATKAEAEKQKAAEAAKAVETEKORAAEATKVAEA
HRAQVHRH	5	5	3	25 38 25 6.13	0	20196_4	VQGVSERVHRAQVHRHSDRPAEHHRVEARRRPTANRH
IKTKQKQL	5	5	5	34 39 37 1.62	0	42130_3	PLNAIKTKQKQLRCSAAHARODGSGACRAKESTNGPAAVW
RRRDQPQH	5	5	5	42 105 53 26.33	0	33041_4	DCEAAHHTASSHPWTRRRDPQHPPPPHSHSQRIWPRPRSQHNAPTPTPATRRKQLSHPRRGDMNLHCHPALNRVPLTCPGHTRCTKPTEKESRAAHSPL
TPRDRWNQ	5	5	3	25 39 31 5.73	1	29663_3	PLHHAQQQTAAAPQQQKRTPRDRWNQCCRCARTPPRPKQ
KRIKWKDN	5	5	3	57 58 57 0.47	0	3671_5	KRIKWKDNDSVSRISVGEREYWSRLTGGGSGIETEDGTLVFPVETGKKEAPNDKKT
TVPHDVTG	5	5	5	38 58 43 7.18	0	9809_5	TPRWTAAHFATSHPLNDITVPHDVTGENQTRNEGRAKPHWRHGGGGPKSSTHGHNPG
PCSRATP	5	5	1	40 40 40 0.0	0	39410_3	CFVPQSQRWRAWHSHGGAVGAPCSRTATPVGETVKGRHPR
PSFSHRGA	5	5	4	39 79 63.0 14.71	0	30083_6	PSFSHRGATPTPTSPASSPTQRRDNSRRCNTRNDQAAAESRCSRHDQTSQMOEIKEKGKLPQASQHNVEAOPAHK
DDQHLPRR	5	5	2	47 79 63.0 16.0	0	16360_6	AARSSQLALNDQHLPRRRHPANLLGNGDTPICRHGRNTRCERVSATPAALCPPRCERMNVPPLPVYTAGAPPTSA
GLPPRPSR	5	5	2	47 47 47.0 0.0	1	37848_3	AHWFYGRYSKQELPRAAPQANVLPSRGLPPRPSRWGRHPHNTDRGP
KRWVIFGK	5	5	5	36 97 63 21.05	0	44885_7	SLGDMKASRERLVGRWRSEKSEHEYANNPSSVGAFTGWSASSRQHENAGDAAKASAPLAESREKRWVIFGKNGSGDKNAGFSTAGTPSKVN
EGEANRSG	5	5	2	36 38 37.0 1.0	1	18737_4	TQRAKYMGRGQQLMHRHTPHNNTQMEEGEANRSGTGS
RAPEQVVK	5	5	2	60 63 61.5 1.5	0	2354_5	FLYNRPLNSTERTAIKDRKVPKRAPEQVVKTPQVAPAVPAGPAGPAGPAGPAGP
MHRTPTTT	5	5	2	36 44 40.0 4.0	0	3938_3	RRGQOPRMHRTPTTHTDTHSNPRQAHTHSPATPTTHRWKREKO
QRAKDGQR	5	5	2	52 60 56.0 4.0	2	12819_4	GFCDSTVSPTSASRAGHYRKCQAYQRAKDGQRTSASSNDSRSPATLQPPAAEPRQ
GPSSRACR	5	5	2	38 40 39.0 1.0	1	7915_3	GPSSRACRWCASGAVGREHTAPLTRGVQQQTIKEGRVEKK
QRRQANPR	5	5	4	44 57 54.0 5.12	1	2842_5	GAAQRRQANPREWKAAPNARNEADSHPRPLRTHPQLPQPSRTHQGDKTERKSPPHNAR
HTHTHTHT	5	4	1	38 38 38 0.0	0	36326_3	GGGGGFGMRGTFEQPHRVAQOREHTHTHTHTQAQNGRA
LRTHRVNG	5	5	4	33 47 36.0 5.34	0	43402_4	VRCSQSGSCALRTHRVNGWMPRSWGDLQSLSPSPSVARGNRESTASLP
SEPVFEVG	5	5	3	56 81 76 10.8	0	14378_5	QQQPQRRQQQTYQQGHDLTSTLQPLSRGHIPSEPSSEGLNSVSSHSRFSRSSEPVFEVGRTHMRLPERRRNTHSNVGHD
IGNTSGTM	5	5	4	30 40 38.0 3.84	1	14219_4	GHAQHVRIGNTSGTMPTSHREPTANATRGKWWGNQRPSSA
ARSLPR	5	5	3	37 41 40 1.7	0	13290_3	VAARRSCSSRSRSDVTASRLCSTSVFACSTARSCLPRSSLLA
KREELQRH	5	5	5	59 85 62 9.7	0	19014_5	NQTKALLDRKMSEVEHNOEQKREELQRHAQERNEAMLAVERRRNLQSOERIEQQOREQORRENHREEQKLEKDLKEQRAE
PSSRTHRS	5	5	2	21 50 35.5 14.5	0	46799_4	IASEASIPKRHRPPSSRTHRSMDEGFFIQTPATLKLPODEMEEDFSYH
RGCRCAAT	5	5	4	30 46 40.0 6.12	0	27536_1	MARCWCCGRRRLCPGRGRCRCAATASAAGSCRCRRIRAVRSSCGAR
IHYRMCVG	5	5	4	25 40 25.0 6.5	0	26644_1	IAPKAKGKKRKHKKRKLKAIHYRMCVGVGEGASVPH
PNRTPPPR	5	5	1	28 28 28 0.0	0	27136_3	HKSHYNTNLKPLYTNPNTPPPRPNP
QWRSGRDD	5	5	4	33 50 43.0 6.42	0	2155_4	PQQSRSPFLTNGKLQWRSGRRDTSQSNGLSREARNGDAPLLWEAQKQT

RRSAARYE	5	5	5	30 41 37 4.03	0	4412_3	SSSGRRSAARYEREVGAGGAAAGWRWRGSGGVVGGSRAS
AERVAER	5	5	3	45 50 49 2.16	0	15418_5	LRAVDGAGR RADCCAERVA AERRCAVSDGRWSAARCGGGEQRGAGGAV
VRVGRCTH	5	5	4	16 25 19.5 3.39	1	18813_1	DPRVVAAHGARARVRVGRCTHWWGG
ARRTGPSS	5	5	3	23 27 25 1.63	1	3141_3	VQQA RRTGPSSWACRW CAGGAVGRETH
DSSGCRHS	5	5	2	46 85 65.5 19.5	0	26888_4	RTL LPRPRTHTRTCDSSGCRHSEHTAHPPAQPPHPHQHLOHSQQPSHTHGRAAAAQERKDQHGAA PPSPPSSLPSPVVESL
LRSGRLWG	5	5	5	23 41 25 6.57	0	35251_3	QLGF EYYLRSGR LWGP PATRPRRCPSPSSTPSPSEVKTH
SSRNRROP	5	5	3	24 35 31 4.55	1	12699_3	SSTKYDYRG ANHDDHPRTV TSSRNRROPQLCVG
PKGRIKPA	5	5	4	38 87 61.0 20.7	0	19915_4	RQEGHTKTRHAPKGRIKPA NSAHRSNTAGSSRTHAPSSSFLISGLPPHPPTAPHNTRGNQKTA EAATTAKPSPCCADATARKGAPA
LGSRGHQA	5	5	2	36 52 44.0 8.0	2	40016_3	ADPSE RGAAMP RGITGLGSRGHQAALHGPPAAWRSHRGHDCGPPACQL
SSEGRCSR	5	5	5	29 39 38 3.66	0	6201_3	RSSAATSSEGRCSRWSVCSRTPRWTSVPCTTWCSSAAP
INSRGLQS	5	5	4	47 61 52.5 5.31	0	33298_5	TLKSPA AERQVSAVATPINSRGLQSRPGSSNKQQQORPWGNEKGAKAREAEAEKS AFVSQH
SKQSPKEP	5	5	1	74 74 74 0.0	0	34134_3	YGHSPAANTORTRSKQSPKEPHTQKRFLTHPQRSAPHHQEETLVRAPSL LPSLHAHGKKQPPHAPQKNHPNPF
SGVRQRCG	5	5	2	42 45 43.5 1.5	0	39819_4	FSHLFTSDHSRRAQIHAVECWH AQPTHTSQAQSGVRQCGCQQR
ARHAPYAA	5	5	2	35 42 38.5 3.5	3	15593_3	TPHEERAAA RHAPYAAQHTLHSTATGTHRHTALPQRQHTDGR
VLWCGAGG	5	5	4	58 69 64.5 4.06	0	2149_6	TGRVLWCGAGGRCEGAEVLGSRGELPPAPELATLPQDTQGLQGGVSGVEDKLLPASSRPLEEEDDD
TTTTRAP	5	5	2	27 38 32.5 5.5	0	30282_3	TTTTTTTAPEAPSI TTTETPNTTTTTRAPSSIRRIDGSL
APSRLREI	5	5	2	27 30 28.5 1.5	1	24746_3	KPPN TTTTTTQA PSTTTTHAPSRLREIDG
SNPLRCWC	5	5	2	22 36 29.0 7.0	0	37345_4	LLGRPAHALQITSNPLCRWC RPLTKKODRPDLLRN
THTHRERE	5	5	1	47 47 47 0.0	0	37968_1	NSTPRHNSQTATHKQPKQKHACPLFH PPTNVHTHTREREGERYHG
KASVYIDG	5	5	2	27 33 30.0 3.0	0	414_4	MLQKKASVYIDGSLGEEDVPLTGEARLGLVH
VDQNTTGD	5	5	5	58 70 63 4.02	0	1718_5	TNEAKHMNGCGVKTPVDQNTTGDGAVANSKKGASGQOKQHPPKGA EKORTEDDVRSVPQKDVNTQVTT
MANIAEKN	5	5	5	35 56 39 8.68	0	35819_5	ATLAGVNGVLN HKPKTAENNIAMANIAEKNRSESRMKTRDYTHGKQPARFRGTGWET
DAAPHSMR	5	5	4	32 63 48.0 11.76	0	44459_4	TRRPLPHHQPPAEADQPSTTRRHEFTLPSR PQDAAAPHSMRSHAMHQTRGKGESGALPPSS
RWEHRCRP	5	5	3	39 45 41 2.49	1	38652_4	VRSSRVALQRQGCVTPRWEHRCRPASHAQEALLPLKKEKNQQPQ
EPOVKIAP	5	5	3	37 52 47 6.24	1	12908_5	PLNSTEMGA IKDRKVPVKRAPEQVKIAPKPAAPVPAVPAGNEGMEKEGD
PRRPAVRV	5	5	5	22 31 22 4.41	0	42437_3	VRQEVSLHGAERRARGVQRRPAVRVKQTE
RPPPTRTA	5	5	2	52 80 66.0 14.0	0	47945_4	ASTRRPGSPSNKTPRSPSSTKSRSQAQORSPTRSQPHAQTTACTAGSRCQQAPPRSRRWQQRHPPTRTAQAPSAPT
PTRTAPAP	5	5	4	32 65 44.5 12.97	0	30184_3	TRSQP HARTPHACTAGSRCQQTPPRSRHWWQHRPPTRTAPAPSAPTARPTALRATTAPRRQQT
VQSHGQRD	5	5	2	30 48 39.0 9.0	1	9881_3	VSNHPWKR RRDLSPEVQSHGQRDTTTAPLSTHNAHRASHTSHT
NGAPIEDG	5	5	5	49 64 63 5.75	0	22626_5	PPLEELRAANGAPIEDGFDAYDRREDDRAARRERVRV RGLNHPRGKPRQNTILKLDSDDEEK

[illegible]