UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE	
FICHA DE EXPECTATIVA DE RESPOSTA DA PROVA ESCRITA	
Edital nº:	035/2017
Carreira:	(x) MAGISTÉRIO SUPERIOR ( ) MAGISTÉRIO EBTT
Unidade Acadêmica:	INSTITUTO METRÓPOLE DIGITAL
Área de Conhecimento:	BIOINFORMÁTICA

CRITÉRIOS DE AVALIAÇÃO PARA TODAS AS QUESTÕES DISCURSIVAS

- Clareza e propriedade no uso da linguagem;
- Coerência e coesão textual;
- Domínio dos conteúdos, evidenciando a compreensão dos temas objeto da prova;
- Domínio e precisão no uso de conceitos;
- Coerência no desenvolvimento das ideias e capacidade argumentativa.

QUESTÃO 1: Considerando dados de expressão gênica, existem várias estratégias de análise que podem ser empregadas para diferentes propósitos. Proponha uma estratégia de análise e apresente como ela opera para se alcançar os seguintes objetivos:

1.1) Seleção de genes significativamente regulados (*Valor: 0,00 a 1,00 ponto*).

1.2) Identificação de genes coexpressos (*Valor: 0,00 a 1,00 ponto*).

1.3) Predição do tipo de câncer a partir do padrão de expressão gênica de um novo paciente, dispondo de um número suficiente de amostras devidamente rotuladas do perfil de expressão gênica de pacientes sem câncer e de pacientes com diversos tipos de câncer *(Valor: 0,00 a 1,00 ponto).*

1.1) Nesta questão, o candidato deve comentar sobre hipótese nula e sobre testes estatísticos como T-test e chi-quadrado, explicando como eles operam.

1.2) Nesta questão, o candidato deve comentar sobre medidas de correlação e técnicas de agrupamento, explicando como elas operam.

1.3) Nesta questão, o candidato deve comentar sobre o tipo de dados que devem estar disponíveis e propor técnicas de reconhecimento de padrões e aprendizado de máquina, como k-NN, regressão logística e redes neurais, explicando como elas operam.

QUESTÃO 2: Considerando a existência de sequências de DNA de várias espécies, com a disponibilidade de múltiplas sequências por espécie, procure responder as questões a seguir:

2.1) Que etapas de pré-processamento das sequências de DNA você proporia para se chegar a uma matriz de distâncias par-a-par entre as espécies? (*Valor: 0,00 a 1,00 ponto*).

2.2) Dadas várias propostas de árvores filogenéticas candidatas a descrever o relacionamento filogenético entre as espécies em estudo, com comprimento de cada ramo definido de acordo com o número de eventos evolutivos esperado entre os seus nós, e supondo que as etapas de pré-processamento das sequências de DNA já foram realizadas e que já se chegou à matriz de distâncias par-a-par para as espécies em estudo (a

partir dessas sequências de DNA), proponha três critérios de desempenho que podem ser considerados independentemente para se medir a qualidade de cada proposta de árvore filogenética. (*Valor: 0,00 a 1,0 ponto*).

2.3) Apresente os principais desafios e dê exemplo de três aplicações práticas da geração massiva de dados a partir de técnicas de sequenciamento de última geração (*Next Generation Sequencing – NGS*). (*Valor: 0,00 a 1,0 ponto*).

2.1) Nesta questão, o candidato deve descrever conceitos e técnicas para o alinhamento de sequências de DNA, possibilitando a comparação direta das sequências. Deve incluir também noções de distância entre grupos de dados.

2.2) Nesta questão, o candidato deve propor máxima verossimilhança (requer um modelo evolutivo), máxima parcimônia (número de eventos evolutivos da árvore, ou seja, somatória dos comprimentos dos ramos) e mínima soma dos erros quadráticos obtidos entre os elementos correspondentes à matriz de distâncias extraída da sequência de DNA e a matriz de distância produzida pela árvore filogenética candidata.

2.3) Desafios: treinamento de profissionais para a realização de análises centradas em big data e questões computacionais como escalabilidade e big data mining, apresentando cenários concretos e atuais. Cabe comentar sobre algumas dimensões envolvidas, como 3 bilhões de bases no genoma humano, ~22.000 genes, com um sequenciamento completo do genoma humano podendo se dar em um dia. Aplicações em potencial, derivadas por exemplo da possibilidade de capturar um espectro mais amplo de mutações: (1) Estimação do risco de desenvolver certas doenças; (2) Apoio à indicação terapêutica; (3) Testes pré-natais. Exemplos de tecnologias: Illumina (Solexa) sequencing; Roche 454 sequencing; Ion torrent: Proton / PGM sequencing; SOLiD sequencing.

<u>QUESTÃO 3</u>: Defina de forma abrangente os seguintes conceitos associados a Big Data:

3.1) Computação em nuvem. (*Valor: 0,00 a 1,00 ponto*).
3.2) MapReduce e Hadoop. (*Valor: 0,00 a 1,00 ponto*).
3.3) Indicadores de desempenho para Big Data. (*Valor: 0,00 a 1,00 ponto*).
3.4) Bases de dados de última geração. (*Valor: 0,00 a 1,00 ponto*).

3.1) Nesta questão, o candidato deve apresentar uma definição conceitualmente correta e abrangente, envolvendo conceitos como independência de plataforma (acesso de qualquer dispositivo físico, rodando qualquer sistema operacional), mobilidade (acesso em qualquer lugar), escalabilidade (recursos "ilimitados"), disponibilidade (acesso a qualquer momento), SaaS e pay per use (evita investimento em recursos que vão ficar ociosos). Associar esses conceitos com vantagens comerciais para empresas e grupos de pesquisa. Exemplos práticos envolvendo provedores específicos (como Google Docs e Dropbox) podem contribuir para enriquecer a resposta.

3.2) Nesta questão, o candidato deve comentar sobre computação paralela e de alto desempenho, tráfego de rede, redundância, tolerância a falhas. MapReduce permite o processamento e a geração de grandes volumes de dados em sistemas computacionais paralelos e distribuídos. Comentar sobre as operações de mapeamento e redução. Uma implementação popular e de código aberto para MapReduce é o Apache Hadoop, que viabiliza a implementação de redes de muitos computadores voltadas para suportar grandes massas de dados e grande volume de computação. O núcleo do Apache Hadoop consiste em uma parte de armazenamento, conhecida como Hadoop Distributed File System (HDFS), e uma parte de processamento, que implementa um modelo de programação MapReduce. Os componentes MapReduce e HDFS do Apache Hadoop foram inspirados pelos trabalhos publicados pelo Google sobre os seus próprios sistemas de arquivo e MapReduce.

3.3) As aplicações bem-sucedidas de big data são aquelas que, de partida, fazem as perguntas certas para se revelar o que se quer efetivamente extrair dos dados. Índices de desempenho devem mensurar / quantificar a eficiência no tratamento de dados de grande volume, não-estruturados e de natureza variada, que fluem em grande velocidade e que podem apresentar ruído e incertezas. Portanto, esses índices devem indicar quando se consegue processar mais dados com os mesmos recursos computacionais, ou então quando se consegue extrair mais conhecimento funcional

da mesma massa de dados ou então o mesmo nível de conhecimento de dados menos estruturados e/ou mais ruidosos. Em suma, o indicador de desempenho deve identificar sistemas de big data capazes de fornecer respostas mais rápidas e mais assertivas, o que implica estimar a relação custo / benefício (investimento na implantação X retorno na operação). Exemplos específicos podem ser utilizados, como em comércio eletrônico, medicina de precisão, bioinformática, smart grid e redes sociais.

3.4) Nesta questão, o candidato deve comentar acerca das novas tecnologias que surgiram após um longo período de predomínio de bases de dados relacionais e representação tabular, gerenciadas por uma linguagem denominada Linguagem de Consulta Estruturada (Structured Query Language, SQL) e produzindo Sistemas de Gestão de Base de dados Relacional (Relational Database Management Systems, RDBMS). Bases de dados não-relacionais (NoSQL) são utilizadas em big data e na web de tempo real, particularmente por empresas como Google, Amazon e Facebook. A principal motivação é trazer escalabilidade na presença de computação distribuída (múltiplos servidores), recorrendo, por exemplo, a representações por valores-chave e grafos, além ou em lugar da representação tabular. Há buscas que não são definidas de forma eficiente em bases relacionais, como web search, onde os dados são semi-estruturados (páginas web), a busca é por palavras-chave e a resposta é uma lista ordenada de itens. Comentar sobre a impossibilidade de preservar transações ACID: Atomicidade, Consistência, Isolamento e Durabilidade. Exemplos de bases de dados NoSQL populares: MongoDB, Couchbase, Riak, Memcached, Redis, CouchDB, Hazelcast, Apache Cassandra, and HBase. Comentar sobre Teorema CAO ou Teorema Brewer: onde é impossível para um sistema distribuído fornecer simultaneamente consistência, disponibilidade, e garantia de tolerância a partição, satisfazendo a apenas duas dessas garantias ao mesmo tempo, mas não todas as três. O NewSQL é uma classe de bancos de dados relacionais modernos que visa fornecer o mesmo desempenho escalável de sistemas NoSQL para cargas de trabalho de processamento de transações on-line (leitura-gravação) enquanto ainda usa SQL e mantém as garantias ACID de um sistema de banco de dados tradicional. Tais bancos de dados incluem o Google F1 / Spanner, o Citus, o CockroachDB, o TiDB, o ScaleBase, o MemSQL, o NuoDB e o VoltDB.

Assinatura dos Membros da Comissão	1º membro (Presidente):  2º membro: Thaís Gaudencio do Rêgo 3º membro: Fernando José Von Zuben