

A estrutura matemática do DNA

Pesquisadores brasileiros mostram por meio de equações que o código genético é similar ao funcionamento do sistema digital

Marcos de Oliveira

Um grupo de pesquisadores brasileiros das universidades Estadual de Campinas (Unicamp) e de São Paulo (USP) está mostrando em artigos científicos que sequências genéticas podem ter uma estrutura matemática semelhante aos Códigos Corretores de Erros (ECC, sigla de *error-correcting codes*) utilizados tanto no sistema de transmissão como de gravação digital. Os ECCs são um conjunto de comandos embutidos em *softwares* instalados nos *chips* de computadores, em equipamentos de telecomunicações, televisores e *smartphones* para corrigir informações digitais com defeitos ao longo de uma conversa telefônica ou, por exemplo, no armazenamento de dados no disco rígido de um computador.

A mesma lógica matemática, de acordo com os pesquisadores, está presente na formação do DNA – o ácido desoxirribonucleico que carrega nas células os genes

e todas as instruções para a formação e a manutenção de um ser vivo. No estudo, eles comparam as equações algébricas de um código corretor de erros com certas sequências do DNA, atribuindo uma lógica aos nucleotídeos que formam o genoma – timina (T), guanina (G), citosina (C) e adenina (A) – e descobriram que há padrões ligando o nucleotídeo a um número. Assim, dependendo do tipo de sequência, o A é representado pelo 0, o C é 2, o G, 1 e o T, 3, por exemplo. Na linguagem digital, formada de *bits*, as informações são traduzidas em 0 ou 1. “Mostramos que o DNA tem sequências que seguem estruturas matemáticas e as mesmas regras da comunicação digital”, conta Márcio de Castro Silva Filho, do Departamento de Genética da Escola Superior de Agricultura Luiz de Queiroz (Esalq), da USP. “A sequência de DNA não é aleatória, segue um padrão”, diz Márcio.



ILUSTRAÇÃO SANDRO CASTELLI



O mais recente estudo do grupo foi publicado em julho na revista *Scientific Reports*, da mesma editora da *Nature*. Na introdução, eles escreveram que os sistemas de comunicação biológica e digital têm semelhanças em relação aos procedimentos correspondentes utilizados para transmitir a informação de um ponto para outro. De acordo com os pesquisadores, a informação contida no DNA é copiada (transcrita) na forma de RNA que irá orientar a ordenação dos aminoácidos nas proteínas necessárias ao funcionamento da célula com uma lógica matemática. No estudo, eles apresentam uma ferramenta computacional para compreender a via evolutiva do código genético ao analisar, por exemplo, a *Arabidopsis thaliana*, planta-padrão de estudos genéticos, e a formação dos nucleotídeos em agrupamentos de três letras chamados de códons. Em casos raros, esse agrupamento biológico – TGA,

por exemplo – apresentou diferenças que não se encaixavam nos resultados apresentados pelo ECC.

Ao apresentar o problema no Congresso Brasileiro de Genética em 2011, Márcio ouviu uma pergunta do biólogo Everaldo Barros, da Universidade Católica de Brasília, que o ajudou a encontrar um caminho. Barros queria saber se aquela alteração em um códon no DNA da batata-doce (*Ipomoea batatas*) não se referia a um código ancestral. Márcio e o engenheiro eletrônico Reginaldo Palazzo Júnior, da Faculdade de Engenharia Elétrica e de Computação (FEEC) da Unicamp, outro coordenador do grupo, empenharam-se em achar a resposta. Junto com as doutorandas Luzinete Cristina Bonani Faria e Andréa Santos Leite da Rocha, eles mostraram que a diferença detectada entre a sequência derivada do código de erros para a biológica é uma mutação que não bate com as equações

matemáticas do genoma primordial da batata-doce, encontrado em sequências de organismos mais antigos como as algas *Prymnesophytes* ou variantes mitocondriais ancestrais do código genético. A mitocôndria é uma organela da célula que guarda resquícios de material genético mais remoto. Assim, apenas o DNA mais antigo se encaixa na equação.

“A sequência do gene que codifica a subunidade delta da proteína F1-ATPase da batata-doce apresenta o códon TGG que codifica o aminoácido triptofano. Entretanto, a sequência gerada pelo código matemático para o códon triptofano era TGA, o que introduziria um *stop* na síntese da proteína, inviabilizando a sua função. A princípio, essa alteração gerada pelo código matemático estaria errada”, diz Márcio. Quando verificamos o aminoácido triptofano que está ali ancestralmente, codificado pelo códon TGA, a conta fechou, então compreendemos que ali aconteceu uma mutação”, diz Reginaldo. Esse tipo de mutação já era conhecido por meio do processo bioquímico, mas nunca havia sido identificado por processo matemático.

Os pesquisadores fazem agora um estudo filogenético para saber mais sobre a evolução das espécies do ponto de vista matemático e biológico. Eles analisam sequências genéticas para verificar se as mutações encontradas apresentam nos indivíduos características relevantes para a funcionalidade da espécie. Os estudos atuais estão sendo feitos em genomas de

No futuro, o estudo por meio de estruturas matemáticas de genes ligados a doenças poderá levar à correção de problemas como o diabetes

E	L	P
GAG	CTT	OCT
101	233	223
101	333	223
GAG	TTT	CCT
E	F	P
L	G	P
TTG	GGA	CCA
331	110	220
331	110	220
TTG	GGA	CCA
L	G	P
D	A	C
GAT	GCA	TGC
103	120	312
103	120	312
GAT	GCA	TGC
D	A	C
T	I	T
ACA	ATC	ACC
020	032	022
020	032	022
ACA	ATC	ACC
T	I	T
V	L	D
GTT	CTT	GAC
133	233	102
133	233	102
GTT	CTT	GAC
V	L	D

plantas e animais para confirmar se de fato o modelo matemático tem uma relação estrita com o que ocorre na biologia.

A descoberta levou o grupo a depositar uma patente internacional do modelo de uso do sistema desenvolvido por eles, que já foi concedida nos Estados Unidos. “Essa estrutura matemática poderá ser importante na área de engenharia de proteínas para a elaboração de organismos geneticamente modificados, novos medicamentos, vacinas e alterar a sequência de DNA em futuros sistemas de terapia gênica, ou, ainda, produzir e descobrir novas proteínas a partir do código matemático”, explica Márcio, um engenheiro agrônomo com mestrado e doutorado em genética e biologia molecular e especialista em transporte de proteínas.

Seria possível também, no futuro, em um tratamento contra o diabetes, por exemplo, estudar os genes ligados à doença por meio de uma estrutura matemática e corrigi-los para que o problema desapareça. Márcio prevê que a indústria farmacêutica terá grande benefício com essa nova forma de ver o DNA porque tanto o entendimento das doenças como a formulação de medicamentos com alvo mais específico de ser atingido estarão facilitados com o uso do código matemático.

ALTERAÇÕES SEQUENCIAIS

Entre matemáticos e cientistas da computação, o código utilizado pelos pesquisadores brasileiros é conhecido pelas letras BCH, iniciais do francês Alexis Hocquenghem e dos indianos Raj Chandra Bose e Dwijendra Kumar Ray-Chaudhuri, que o descobriram entre 1959 e 1960. O BCH é apenas um dos códigos corretores de erros existentes. Utilizando esse código, biólogos, bioquímicos e farmacêuticos, possivelmente com a colaboração de matemáticos, poderão fazer análises preliminares com as sequências no computador para testar a alteração de aminoácidos, proteínas e mutações e só então ir para o laboratório verificar se o resultado está correto. “A existência de uma estrutura matemática em sequências de DNA implica uma complexidade computacional enorme, porém factível para a realização de análises e previsões de mutações”, diz Reginaldo, que é engenheiro eletrônico e atua nas áreas das teorias da informação e codificação. Atualmente, esse processo de alteração para produção de um organismo geneticamente modificado ou de



um medicamento é realizado por meio de extensivos testes laboratoriais. A função do código matemático nos processos biotecnológicos será de minimizar a ocorrência de erros no núcleo da célula, depois da transcrição gênica do DNA para o RNA, o ácido ribonucleico que dirige a síntese da proteína nos ribossomos.

A possibilidade da associação de códigos de correção de erros com sequências de DNA não é nova. Um dos principais estudiosos do assunto é o professor Hubert Yockey, que atua na área desde a década de 1980 na Universidade da Califórnia em Berkeley, nos Estados Unidos. Outro pesquisador da área é Gérard Battail, professor aposentado da Escola Nacional Superior de Telecomunicações, na França, que publicou vários artigos propondo a relação entre código de correção de erros e genoma. Eles têm demonstrado o processo e levantado hipóteses, mas não apresentaram as efetivas relações matemáticas com o DNA. Os brasileiros conseguiram estabelecer essa relação nas sequências genéticas produtoras de proteínas. “Ao conhecermos a estrutura matemática do gene que codifica a proteína, é possível alterar a ordem das bases e também corrigir as mutações ou erros que possam acontecer para ela voltar à condição original de uma proteína”, diz Márcio.

O estudo inicial surgiu com Reginaldo, que indicou às duas doutorandas, em 2008, o desafio de modelar a transmissão da informação, no caso as proteínas, entre o núcleo celular e a mitocôndria. Para isso, Luzinete e Andréa procuraram então Márcio de Castro, na Esalq. O diálogo foi estabelecido e as duas passaram a testar alguns modelos matemáticos de sistemas de comunicação com o objetivo de encontrar um que se adequasse ao modelo biológico. Depois de alguns meses, elas mostraram o resultado para Márcio, que, no início, pensou haver apenas uma coincidência entre as sequências geradas pelo ECC e a biológica em relação aos aminoácidos. Com o avanço dos estudos, foram realizados levantamentos de sequências de DNA de diferentes seres vivos e o resultado se manteve independentemente da espécie. A descoberta teve também a participação do então doutorando João Henrique Kleinschmidt, engenheiro da computação que hoje é professor na Universidade Federal do ABC (UFABC), e mais

L	K	G
TTG	AAA	GGA
331	000	110
331	000	110
TTG	AAA	GGA
L	K	G
E	L	P
GAG	CTT	CCT
101	233	223
101	333	223
GAG	TTT	CCT
E	F	P
L	G	P
TTG	GGA	CCA
331	110	220
331	110	220
TTG	GGA	CCA
L	G	P

Nesta página e na outra, as letras e número em vermelho indicam mutação na sequência genética

recentemente da bióloga Larissa Spoladore, doutoranda da Esalq, e do biólogo Marcelo Brandão, professor da Unicamp.

Em 2009, Márcio, Reginaldo, Luzinete e Andréa submeteram um artigo à revista *Electronic Letters*, que foi publicado na edição de fevereiro de 2010 (ver Pesquisa FAPESP nº 178). “Agora com a publicação na *Scientific Reports*, acreditamos que a comunidade mundial da área das ciências biológicas poderá se interessar mais”, diz Márcio. “Não existe outro grupo fazendo pesquisa nesse sentido segundo a literatura aberta, até onde sabemos, porque pode ter alguém da indústria farmacêutica, de forma fechada, desenvolvendo algo nesse sentido.”

“Como várias outras descobertas científicas, essa deverá ter um caminho longo para ser aceita e utilizada. Eles deram um salto que, sem nenhuma dúvida, é uma quebra de paradigma”, diz o biólogo Rogério Margis, professor do Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul (UFRGS). “Imagino que existirão novos desafios com a descoberta desse padrão que transcende a sequência linear das bases e adiciona mais uma camada de complexidade e de padrões de código na molécula de DNA. Para expandir esse tipo de análise será necessário ter uma grande infraestrutura computacional”, comenta Rogério.



“Até agora os estudos deles não tiveram o impacto e a repercussão esperada na comunidade científica. Um problema é que o estudo, embora único, engloba áreas distintas como biologia e matemática, que pouco conversam”, diz.

“Já apresentei os estudos em eventos no exterior, mas acho que há uma certa desconfiança por várias razões. O assunto é extremamente complexo, poucas pessoas conseguem transitar nas duas áreas, da genética e dos códigos corretores de erros, o grupo é de brasileiros e o trabalho de 2010 foi divulgado em uma revista da área de engenharia elétrica”, explica Márcio. O interesse maior pelos estudos, segundo ele, deve partir do pessoal da biologia molecular e da biotecnologia. Do lado da matemática, seriam os grupos de teoria da informação e comunicação. Mas isso só vai acontecer se houver uma integração multidisciplinar, como aconteceu no caso da descoberta. ■

Projetos

1. Código matemático de geração e decodificação de sequência de DNA e proteínas: utilização na identificação de ligantes e receptores (nº 2008/04992-0); **Modalidade** Programa de Apoio à Propriedade Intelectual (Papi); **Pesquisador responsável** Márcio de Castro Silva Filho (USP); **Investimento** R\$ 13.200,00 e US\$ 20.000,00.
2. Herbivory and intracellular transport of proteins (nº 2008/52067-3); **Modalidade** Projeto Temático; **Pesquisador responsável** Márcio de Castro Silva Filho (USP); **Investimento** R\$ 1.392.217,77 e US\$ 169.187,06.
3. Biologia de sistemas aplicada à agricultura: análise de transcriptomas e interactomas (nº 2011/00417-3); **Modalidade** Programa Jovens Pesquisadores em Centros Emergentes; **Pesquisador responsável** Marcelo Mendes Brandão (Unicamp); **Investimento** R\$ 199.169,39 e US\$ 3.846,15.

Artigos científicos

- BRANDÃO, M. M., et al. Ancient DNA sequence revealed by error-correcting codes. **Scientific Reports**. v. 5, n. 12051, jul. 2015.
- FARIA, L. C. B., et al. Transmission of intra-cellular genetic information: A system proposal. **Journal of Theoretical Biology**. v. 358, p. 208-31, out 2014.
- FARIA, L. C. B., et al. Is a Genome a Codeword of an Error-Correcting Code? **PLoS ONE**. v. 7, n. 5, e 36644, mai. 2012.
- FARIA, L. C. B. et al. DNA sequences generated by BCH codes over GF(4). **Electronics Letters**. v. 46, n. 3, p. 202-3, fev 2010.