USP ESALQ – ASSESSORIA DE COMUNICAÇÃO
Veículo: Pesquisa FAPESP
Data: 12/2010
Caderno / Página: Matemática / 54 a 57
Assunto: Equações da vida e códigos seqüências de DNA

Equações da vida

A mesma estrutura de códigos une sequências de DNA e comunicação digital

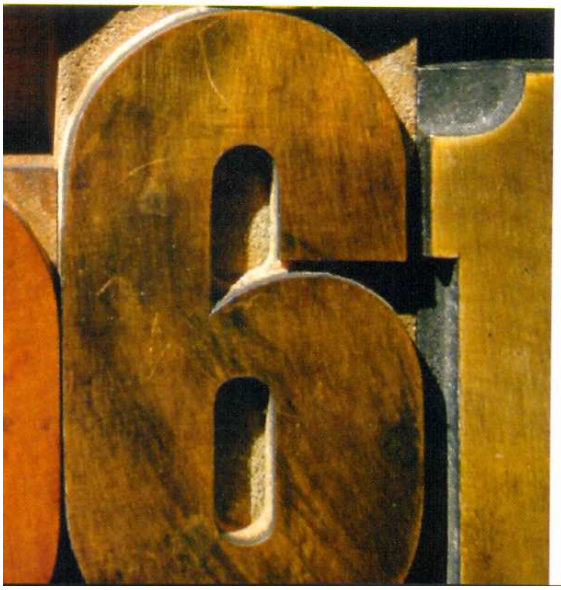
MARCOS DE OLIVEIRA

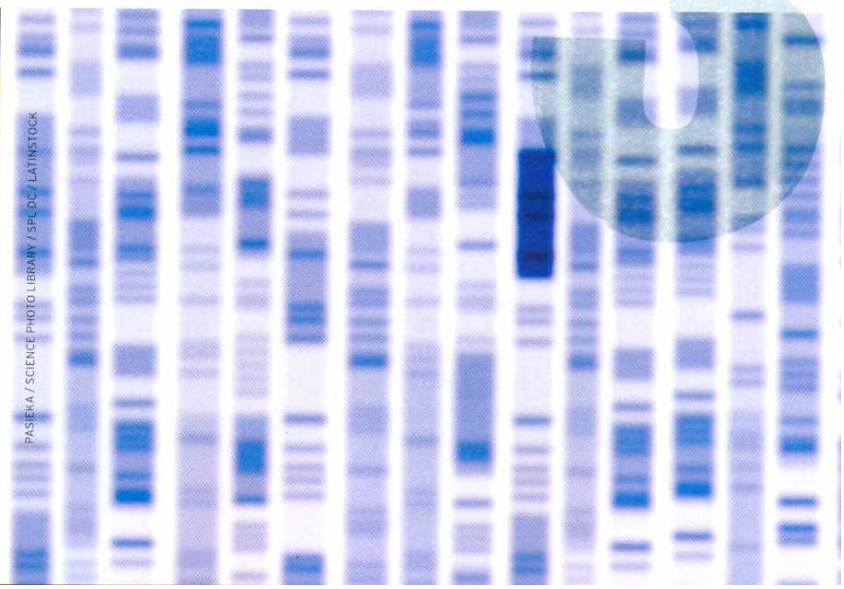
Explicar os fenômenos da natureza com equações matemáticas é uma tarefa rotineira e incorporada aos estudos da física, da química e da própria matemática. A biologia tem uma tradição menor nesse sentido. Essa relação é perseguida por vários grupos de estudo na Europa e nos Estados Unidos que buscam uma vinculação dos genomas de seres vivos com estruturas matemáticas para tentar compreender melhor a formação da vida no planeta. Mas a primazia de encontrar tal vínculo coube a um grupo de pesquisadores da Universidade Estadual de Campinas (Unicamp) e da Universidade de São Paulo (USP) que encontraram uma relação matemática entre um código numérico e a sequência do DNA, a sigla em inglês do ácido desoxirribonucleico que carrega os genes dentro das células. Outros pesquisadores já haviam sugerido essa relação, mas não conseguiram provar. Os brasileiros descobriram que as bases nitrogenadas timina (T), guanina (G), citosina (C) e adenina (A) se organizam segundo uma lógica numérica. "A distribuição dessas bases possui um código matemático que prevaleceu ao longo da evolução dos seres vivos", diz o professor Márcio de Castro Silva Filho, da Escola Superior de Agricultura Luiz de Queiroz, da USP. "Descobrimos que uma proteína ao perder a função biológica devido a


ELIZABETH PAPADOPOULOS / GETTY IMAGES

uma mutação, por exemplo, deixa de ser representada por uma estrutura matemática", diz Silva Filho, um dos coordenadores do grupo.

Os pesquisadores não desenvolveram um código novo para explicar a sequência do DNA. Eles verificaram que existe uma relação entre certas sequências de DNA com o código corretor de erros (ECC, sigla de *error-correcting code*), que são equações matemáticas utilizadas em todo processo digital, usado em sistemas de comunicação e de telecomunicações, em memórias de computador e memórias *flash* de *pen-drives* para corrigir ruídos ou defeitos que surjam nas transmissões. O código também é conhecido pelas letras BCH, que são as iniciais de seus descobridores – os indianos Raj Chandra Bose e Dwijendra Kumar Ray-Chaudhuri e o francês Alexis Hocquenghem –, e não apenas identifica o erro mas também faz a correção. A atribuição da associação de códigos de correção de erros com sequências de DNA não é nova. É objeto de pesquisa desde a década de 1980 e um dos principais estudiosos é o professor Hubert Yockey, que trabalhou na Universidade da Califórnia em Berkeley, nos Estados Unidos, e publicou dois livros: *Information theory and molecular biology*, em 1992, e *Information theory, evolution, and the origin of life*, em 2005, ambos editados pela Cambrigde University Press. Outro pesquisador da área é Gérard Battail, professor aposentado da Escola Nacional Superior de Telecomunicações, da França, que tem escrito artigos propondo a relação entre código corretor de erros e o genoma. Eles têm demonstrado o processo e levantado hipóteses, mas não apresentaram as relações matemáticas com o DNA. Os brasileiros conseguiram estabelecer essa relação nas sequências do ácido ribonucleico mensageiro (RNAm) que geram as proteínas.





PASIEKA / SCIENCE PHOTO LIBRARY / SPL DC / LATINSTOCK

"Ao conhecermos a estrutura matemática da proteína é possível alterar a ordem das bases e também corrigir as mutações ou erros que possam acontecer para voltar à condição normal de uma proteína", diz o professor.

Problema molecular - A capacidade de corrigir uma mutação ou um erro celular poderia, por exemplo, no futuro utilizar uma solução matemática para atuar sobre a falta de produção de insulina pelas células do pâncreas, corrigindo erros em um gene específico. "Seria possível identificar a estrutura matemática das mutações e onde elas ocorreram e talvez corrigir esse problema molecular para o organismo voltar a produzir insulina, revertendo as estruturas anteriores. Outra possibilidade seria fabricar proteínas a partir do código matemático ou ainda encontrar proteínas não conhecidas existentes nas células", diz o professor Reginaldo Palazzo Júnior, da Faculdade de Engenharia Elétrica e de Computação (Feec) da Unicamp, outro coordenador do grupo. "A correção ou a forma de reverter o erro nas células acontece da mesma forma que num disco rígido (HD), que tem um setor danificado e o ECC reconstitui as informações."

Com tantas possibilidades de uso na indústria, além do significado científico importante da descoberta, os pesquisadores resolveram, antes de publicar a novidade em periódicos científicos, depositar uma patente internacional pelo Tratado de Cooperação em Patentes (PCT, na sigla em inglês), em vários países, e outra nos Estados Unidos, com financiamento da FAPESP e gerenciamento da Agência de Inovação da Unicamp e da Agência USP de Inovação. Laboratórios do mundo poderão usar, se licenciarem a patente, as estruturas matemáticas descobertas pelo grupo, possivelmente na forma de um *software*, para testar proteínas em um amplo leque de produtos. "Essas informações são importantes para desenvolver vacinas, medicamentos ou proteínas para elaboração de queijos e amaciantes de roupa, por exemplo", diz o professor Silva Filho. Hoje se faz uma alteração na sequência de DNA que codifica uma proteína e depois são feitos os testes em laboratório para verificar a eficácia da reação num experimento de tentativa e erro. Com as equações matemáticas será possível testar a afinidade e a estabilidade da proteína em um trabalho preliminar de forma a verificar mutações e, posteriormente, testá-las a partir de experimentos de laboratório para confirmar se a mutação na sequência de DNA deu o resultado esperado. "Se a estrutura matemática não se mantiver, a alteração não vai ser efetivada e não produzirá os resultados esperados."

A descoberta da existência de um código matemático que transcreve a sequência de DNA aconteceu quase por acaso e começou com o professor Palazzo, que lançou um desafiante objetivo a duas alunas de doutorado, Luzinete Cristina Bonani Faria e Andrea Santos Leite da Rocha, orientadas por ele na Feec e oriundas da graduação da Pontifícia Universidade Católica de Campinas (Puccamp), com mestrado na Unicamp. Elas deveriam procurar as informações que transitam dentro de uma célula. "Dentro da mitocôndria, um órgão responsável pela respiração celular, existem moléculas de DNA para sintetizar certas substâncias, mas ela não tem todas as proteínas e precisa solicitar proteínas extras produzidas por genes localizados no núcleo de modo a realizar as funções na organela. Nesse

O PROJETO

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)

COORDENADOR
Márcio de Castro Silva Filho - USP

INVESTIMENTO
R$ 13.200,00 e US$ 20.000,00 (FAPESP)

caso, para os matemáticos, a proteína é informação e existe um código padrão para transmiti-la", explica o professor Palazzo. O modelo apresentado pelos pesquisadores brasileiros se ajusta a qualquer sequência de DNA produtora de proteínas dentro da célula.

Palazzo é um especialista na chamada teoria matemática da comunicação, área de estudo que pesquisa a transmissão de todo tipo de informação e seus códigos. Também chamada de teoria dos códigos, ela analisa as formas de transmissão independentemente do significado. Assim não importa a palavra que está sendo transmitida, mas como ela é enviada de um emissor A para um receptor B, dentro de um contexto matemático. "Essa teoria foi apresentada por Claude Shannon [matemático e engenheiro eletrônico norte-americano] em 1948", lembra Palazzo.

Para o estudo de Andrea e Luzinete, Palazzo sugeriu que elas procurassem os professores da Unicamp, da área da Faculdade de Ciências Médicas (FCM), inicialmente, para encontrar componentes biológicos e se aprofundar no tema. Depois de muita procura, elas ouviram a sugestão do professor Anibal Vercesi, da FCM, para procurarem o professor Márcio de Castro Silva Filho na Esalq. "Fomos conversar com ele e estabelecemos um casamento de interesses", diz Palazzo. "Começamos um diálogo tendo de um lado matemáticos e um engenheiro elétrico e eu, um geneticista especializado em transporte de proteínas", lembra Silva Filho. A primeira amostra de DNA investigada pelas pesquisadoras da Unicamp foi da *Arabidopsis thaliana*, planta da família da mostarda, que serve de modelo para estudos genômicos. A partir daí, elas ficaram trabalhando durante seis meses. "Começaram a testar vários elementos matemáticos para tentar achar alguma sistematização em relação ao genoma", explica Palazzo, que contou também com a colaboração no estudo do engenheiro da computação João Henrique Kleinschmidt, ex-aluno de doutorado e atual professor da Universidade Federal do ABC, em Santo André, na Região Metropolitana de São Paulo. "Um dia elas me chamaram na Unicamp e me mostraram os resultados. Quando percebi o que era fiquei sem fala. Pensei que fosse uma coincidência e passamos a repetir o trabalho usando outros genomas, do homem, de bactérias, fungos e plantas. Descobrimos que é um processo universal", conta Silva Filho.

> A teoria da informação é uma ferramenta adequada para o intercâmbio com a biologia molecular, diz Gérard Battail

Entender a linguagem - No final de 2009, eles submeteram um artigo às revistas *Nature* e *Science*, mas as duas recusaram dizendo que era algo muito específico. "Acredito que eles não entenderam a linguagem matemática do *paper*", diz Silva Filho. "Isso faz parte da dificuldade da conversa entre biólogos, engenheiros, médicos etc.", diz Palazzo. Aí eles resolveram enviar para a revista *Electronics Letters*, que em três semanas aceitou o trabalho e o elegeu o melhor artigo de fevereiro deste ano, colocando-o na capa do mesmo mês. Eles começaram a mostrar o estudo em congressos internacionais de teoria da informação e devem apresentar novos resultados com informações mais detalhadas e com outras ferramentas matemáticas. No artigo da *Electronics Letters*, "DNA sequences generated by BCH codes over GF(4)", ou "Sequências de DNA geradas pelo código BCH sobre GF(4)", eles apresentaram uma parte do trabalho utilizando a estrutura matemática chamada de corpo algébrico de Galois, enquanto novos resultados usam a estrutura de anel de Galois. Em uma simplificação, poderíamos dizer que em relação ao corpo o produto de dois números diferentes de zero resulta em um número diferente de zero, enquanto na estrutura de anel o produto pode ser zero. Para os matemáticos isso faz muita diferença na apresentação dos resultados. Até agora eles apresentaram apenas os resultados em corpo.

O feito dos pesquisadores brasileiros apresenta uma solução importante e uma novidade para a biologia. Ela inicia uma nova fase em que os fenômenos que estuda passam a ser analisados por métodos quantitativos. "Em 1999, a Academia Real da Suécia indicou que um dos avanços da ciência no novo século seria a incorporação de mais matemática aos estudos da biologia", lembra Silva Filho. Mas para isso tanto os pesquisadores brasileiros como Battail e Yockey concordam que é preciso um maior diálogo entre biólogos, matemáticos e engenheiros eletrônicos. "Como engenheiro, eu estou convencido de que a teoria da informação é uma ferramenta adequada para intercâmbio com a biologia molecular", escreveu Battail em uma apresentação do livro de Yockey em 2006. "Ainda estamos distantes de uma interdisciplinaridade que permita a conversa entre áreas para projetos desse tipo. Mas nós já demos um bom passo", diz o professor Palazzo. ■

Artigo científico

FARIA, L.C.B.; ROCHA, A.S.L.; KLEINSCHMIDT, J.H. ; PALAZZO Jr., R.; SILVA FILHO, M.C. DNA sequences generated by BCH codesover GF(4). **Electronics Letters.** v. 46, n. 3, p. 202-03. fev. 2010.