Boletim

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Publicação semestral *outono de 2015*

Estatística em Genética

Informação Editorial

Endereço: Sociedade Portuguesa de Estatística.
Campo Grande. Bloco C6. Piso 4.
1749-016 Lisboa. Portugal.
Telefone: +351.217500120
e-mail: *spe@fc.ul.pt*
URL: http://www.spestatistica.pt
ISSN: 1646-5903
Depósito Legal: 249102/06
Tiragem: 500 exemplares
Execução Gráfica e Impressão: Gráfica Sobreirense
Editor: Fernando Rosado, *fernando.rosado@fc.ul.pt*

WORLDOFSTATISTICS.ORG

O MUNDO DA
ESTATÍSTICA

ORGANIZAÇÃO PARTICIPANTE

World Statistics Day

Service • Professionalism • Integrity

WORLD STATISTICS DAY

20.10.2015

BETTER DATA. BETTER LIVES.

DIA MUNDIAL DA ESTATÍSTICA

20.10.2015

MELHORES DADOS. MELHORES VIDAS.

Editorial

… de novo, a força da Estatística, também nas Sondagens…

1. O **Congresso Internacional ISI 2015**, decorreu no Rio de Janeiro desta vez com a notícia especial da liderança da lusofonia como anunciámos no Boletim SPE primavera 2015. A consulta do programa desse congresso também revela um bom número de participantes portugueses. Foi em Lisboa num congresso recente do ISI em 2007 que, bem perto de nós, também se viu a força da Estatística com o número record de participantes que responderam presente após convite da organização portuguesa desse acontecimento de relevo mundial. Naquele Congresso Mundial, **Explorística**, de novo, recebeu um Prémio como noticiamos neste Boletim.

2. O **XXII Congresso SPE** retomou uma tradição, interrompida em 2014. O sucesso verificado, desde logo pelo dinamismo (uma tradição que se manteve!) e também o elevado número de participantes justificam que, como deve, esta seja uma iniciativa que merece a melhor atenção. O sucesso do XXII Congresso SPE reforça a convicção da sua necessidade. Mesmo com o aumento da oferta de conferências e com a drástica diminuição das bolsas, isto é, com a consequente dificuldade de financiamento agravadas, os estatísticos desejam continuar esta atividade, reconhecidamente, basilar.

Ela é fundamental para a divulgação da Ciência Estatística em Portugal (e no mundo) e também como um estímulo ao seu estudo por parte dos mais jovens investigadores – a jeito de iniciação…

Além de tudo, a não existência de congresso… faz esquecer! O não aparecer pode significar (também) inatividade e menor dinamismo por parte dos investigadores portugueses e não só.

O Congresso SPE é a ação que mais concretiza os objetivos estatutários. O incentivo e a "primeira experiência" para os jovens investigadores, como é testemunhado neste Boletim SPE, é igualmente um ponto do maior interesse a que se deve dar muita atenção. Essa ação ajuda a construir um sentimento de missão cumprida. Estão pois de parabéns a atual Direção bem como a Comissão Organizadora e a Comissão Científica. Como memorial, que já é uma tradição, nesta edição, graças à generosidade das colegas Cristina Miranda e Anabela Rocha, fica também o XXII Congresso registado para a história. Como ponto de programa da atual Direção deseja-se que se repense o assunto. Jovens e seniores, são diversas as vozes que ouvi no congresso e que fazem esse apelo.

3. As sondagens e os estudos de opinião com base nelas, nos últimos tempos, têm sido alvo de bastantes referências nos mais diversos meios de comunicação. Os mais recentes textos surgiram, obviamente, na sequência das eleições legislativas do passado dia 4 de outubro em Portugal.

No entanto, em maio, a propósito de eleições no Reino Unido, comentava-se: "A maioria absoluta dos conservadores ultrapassou em muito as sondagens da véspera. Mas as razões dos desvios são, até ao momento, insondáveis" ou "As razões por detrás do falhanço das sondagens são ainda, em grande parte, desconhecidas". Alguns admitiram que "o falhanço das sondagens nas eleições legislativas britânicas" se deveu ao receio de as pessoas serem vistas como "politicamente incorretas". Outros, usaram "a teoria do silêncio" para justificar que "impulsionados pelo medo de isolamento social" alguns votantes tenderão a ser "menos propensos a exprimir os seus pontos de vista, sempre que acharem que as suas opiniões e ideias são minoritárias". É uma espécie de auto censura, diz-se.

Os motivos (para uma possível explicação do falhanço) são de tal forma desconhecidos que o Conselho de Sondagens Britânico anunciou que seria lançado um inquérito para se chegar às causas dos erros. Mas erros, a havê-los, podem não ter acontecido apenas nestas eleições britânicas. Ao longo dos últimos anos têm-se multiplicado os casos de sondagens pré-eleitorais que falharam o alvo.

Neste ano, também em Israel e na Escócia, a Ciência Estatística foi posta à prova e, nalguns casos, teve de recorrer ao "erro que a protege".

Está (também) por aqui a mão dos "brancosos" de Saramago no *Ensaio sobre a Lucidez*?

A SPE atua em diferentes campos e nos mais diversos domínios, em particular, na área das Sondagens e Estudos de Opinião. No seu património de intervenção científica, a SPE tem estimulado as mais variadas contribuições registadas praticamente em todas as Atas dos Congressos SPE. A primeira edição de livros de minicursos, em 1998, foi *Tópicos de Sondagens* da autoria de Paulo Gomes.

No Boletim SPE primavera de 2011 o tema central foi Sondagens e Censos. Passados alguns anos e perante a oportunidade dos desafios e debates emergentes do que acima referenciei está na hora de revisitar o tema e dedicar uma nova edição do Boletim SPE centrado nas sondagens.

Os meios de comunicação social emitem comentários e análise nos mais variados campos relacionados com os estudos de opinião. Esta incursão, nalguns casos com "discursos" inusitados e atrevidos podem surgir pela ausência de temas que seriam naturais se a política estivesse no centro?

A opinião pública, diz-se, aceita com mais reserva, tem maior desconfiança ou põe mais facilmente em causa uma sondagem favorável ao governo do que uma favorável à oposição. Constata-se que a abstenção diminui quando se deseja que um de dois candidatos não vença e cresce sempre que é mais ou menos irrelevante que ganhe um ou o outro. A abstenção decide? É (mais) um menor!? Vários pormenores ficam assim alinhados para possíveis explicações.

4. A Estatística é a ciência dos dados, também aplicada, porque a pesquisa muitas vezes, visa também uma aplicação. A Estatística é interessante e útil porque fornece estratégias e instrumentos para trabalhar os dados de modo a melhor "entrar" em problemas reais. Dados são números (ou a falta deles) inseridos num determinado contexto ou experiência. Mas, determinar a média de 50 números, é puro cálculo aritmético, não é Estatística. Discernir sobre aquele valor 50 e decidir se temos uma pequena ou grande amostra e, em cada caso, concluir sobre a discrepância de determinado valor (mesmo que usando a média atrás calculada!) já é Estatística. Embora a Estatística se possa considerar como uma ciência matemática, ela não é um ramo da matemática e não deve ser ensinada como tal. Cada vez mais, podemos falar em pensamento estatístico que suporta e se apoia na teoria da decisão.

Ao meditar sobre a investigação e a teoria da decisão tenho referido o lema "*quos fama obscura recondit*". Esta expressão primorosa de Virgílio (Eneida, V, 302) foi glosada, entre muitos outros, por Santo Agostinho (*A Cidade de Deus*, volume I, Livro VII, Capítulo III, p. 611 e seguintes. Serviço de Educação. Fundação Calouste Gulbenkian. 1991). Na dicotomia entre a "razão menor" e uma "razão mais alta" deve o estatístico ter como objetivo (apenas) o conhecimento que lhe permite cobrir as suas necessidades científicas básicas? Em alternativa, esse deve ser um estádio inicial tendo por objeto a sabedoria estatística onde (ainda) admite a (enorme) importância dos "detalhes científicos" daqueles a quem uma obscura fama esconde – chamemos-lhes, por exemplo, *outliers*; que são estimuladores da investigação e podem ser originados pelos valores discordantes de uma amostra – uma minoria. Acima, sobre as Sondagens "aparecem" muitos "detalhes". O seu estudo contribui para afirmar a Estatística. São esses "menores" – aqueles a quem uma obscura fama esconde – que fazem avançar a ciência?! Neles está a força! É a *Força dos Menores* que leva a desconfiar das sondagens?

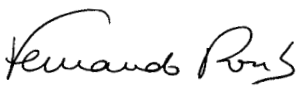
Assim, sem querer(?) quis o acaso – as eleições e as sondagens – que eu desse sequência ao sugerido por um dos *referees* do meu artigo "*Outliers*: The Strength of Minors" publicado em 2014 em *New Advances in Statistical Modeling and Applications* (Pacheco *et al*, Editors) p. 17-27. Springer. *Studies in Theoretical and Applied Statistics*. Aquele, satisfeito com o que leu, sugeria que eu aprofundasse um pouco mais. Aqui está mais um pouco!

Perante as diversas problemáticas e questões enunciadas, o que fazem os estatísticos? O que devem fazer os estatísticos? As respostas, possivelmente, também envolvem as associações científicas – desde a Estatística às Ciências Sociais e Políticas. É fundamental refletir sobre a influência das novas, digamos modernas, variáveis intervenientes na metodologia tradicional. A "formação de opinião" é hoje muito mais rápida? Isto é, as chamadas "redes sociais" têm o poder de, rapidamente, alterar ou manipular as tendências? E, sendo assim, elas podem fazer "flutuar" a opinião pública que, flutuando, não fica estabilizada e muito mais vulnerável. São o mais recente "menor"?

Eu não sou especialista em Sondagens e Estudos de Opinião mas, pelo exposto, sinto uma enorme motivação para revisitar este tema em próximo Boletim SPE.

Fica o desafio aos especialistas! Pela minha parte, como editor, vou iniciar o trabalho de casa.

O tema Central do próximo *Boletim* será *Séries Temporais e suas aplicações.*



Mensagem da Presidente

Caros sócios da SPE

Escrevo este texto no rescaldo do XXII Congresso da Sociedade pelo que, naturalmente começo por fazer uma breve referência a este evento que é sempre marcante para a nossa Sociedade. A realização deste Congresso foi posta em marcha pela Direção anterior da SPE, representada na mesa de abertura do Congresso pelo colega Daniel Paulino e sua concretização deve-se fundamentalmente à Comissão Organizadora. Foi um trabalho extenso e generoso de muitos e por isso quero, em nome da atual Direção, exprimir publicamente a nossa gratidão à Comissão Organizadora pelos esforços que desenvolveu para levar a cabo esta iniciativa, com o sucesso que pudemos testemunhar, apesar de vicissitudes e dificuldades várias que teve de ultrapassar. Um agradecimento especial é devido também à Comissão Científica por todo o apoio prestado às solicitações da Comissão Organizadora. Quero expressar os nossos agradecimentos às diversas entidades que patrocinaram e apoiaram a organização, pelo seu auxílio na concretização de uma iniciativa desta envergadura. A estatística é uma área intimamente ligada aos problemas da sociedade pelo que ter o envolvimento e apoio da sociedade civil é para nós um registo importante. Quero agradecer aos oradores convidados por terem acedido ao convite formulado pela Comissão Organizadora, prestigiando este congresso com as suas apresentações. Em particular aos colegas Maria Antónia Turkman e Daniel Paulino pelo curso sobre Estatística Bayesiana que foi informativo e estimulante. Por último e relembrando o adágio popular que afirma que os últimos são os primeiros, temos de agradecer a todos os participantes o contributo não apenas para o sucesso deste congresso como para a vitalidade da Estatística em Portugal. Este congresso realizou-se após uma quebra estrutural na variável tempo entre 2 congressos, contou com 4 conferências plenárias e um total de 114 comunicações orais e posters, numa grande diversidade de tópicos quer de cariz mais metodológico quer de cariz aplicado. Algumas destas comunicações são de jovens estudantes de mestrado e doutoramento a quem foi concedida, pela Direção, uma bolsa de participação no Congresso com base no mérito dos trabalhos submetidos. Dois momentos importantes deste Congresso para a nossa Sociedade uma vez que a unem e atestam a sua vitalidade, foram: a Sessão dedicada ao Prémio SPE 2015 e a Sessão de atribuição do Prémio Carreira SPE que homenageou a Professora Doutora Maria Antónia Amaral Turkman. A componente social do Congresso, em particular o passeio e o jantar, são sempre momentos altos e a CO deste XXII Congresso respondeu adequadamente às expectativas. Um relato minucioso destes momentos pode ser encontrado nas páginas deste boletim. Os Congressos da SPE têm o sabor de uma reunião de família: é a nossa família científica e muitos encontram aqui os seus avós, pais, irmãos e primos científicos. É certamente esta a explicação para um tão grande número de participantes mesmo sob severas restrições de financiamento. Resta-nos esperar por notícias sobre o próximo Congresso que terá lugar em 2017.
Nestes quase 10 meses de trabalho, a Direção deparou-se com dificuldades relacionadas com parte informática e o secretariado da Sociedade. Estas dificuldades estão quase ultrapassadas mas agradecemos que eventuais erros e omissões na página web sejam reportados. Esta mensagem já vai longa mas não posso deixar de referir os seguintes assuntos. A Explorística ganhou mais um prémio internacional: IASE-ISLP Best Project Award in Statistical Literacy. A versão em Inglês encomendada pela Irlanda está em produção e já foram entregues alguns módulos. Vai estar exposta em Lugo durante o Congresso da SGAPEIO. Os Prémios Estatístico Júnior 2015 foram entregues numa sessão pública na FNAC de Santa Catarina no Porto. Contamos com a presença de premiados e famílias num total de cerca de 70 pessoas. Foram atribuídas pela primeira vez em 2015 bolsas de participação no Congresso a estudantes de Mestrado e Doutoramento. Foi deveras gratificante verificar a qualidade dos trabalhos submetidos. Estão em curso diversas iniciativas das quais daremos conta no site da SPE, por correio eletrónico e em próximos Boletins. Especial atenção é devida à organização conjunta SPE-SGAPEIO do *II Encontro Luso-Galaico em Biometria com aplicações à saúde, ambiente e ecologia* que terá lugar em Santiago de Compostela de 30 de Junho a 2 de Julho 2016.

Até breve,

Porto, 13 de Outubro de 2015
Cordiais saudações
Maria Eduarda Silva

Notícias

• XXII Congresso SPE

O arranque dos trabalhos

Com alguma informação *a priori*, criámos expetativas elevadas para mais uma edição do nosso congresso da SPE. A página na *net* indiciava uma série de acontecimentos multivariados que atraíram a comunidade dos estatísticos portugueses, brasileiros, cabo-verdianos e espanhóis, refletindo-se num elevado número de cerca de 180 participantes.
Fomos recebidos com um sorriso acolhedor e uma vista magnífica sobre a Ria Formosa, logo na manhã do dia 7. Tivemos a certeza de que a viagem não tinha sido em vão, ao sermos brindados com um curso introdutório de Estatística Bayesiana Computacional, com oradores "estrelados": pela manhã, com uma energia despertadora, o Prof. Paulino falou sobre formulação e análise dos modelos bayesianos; à tarde, com graciosidade, a Prof.ª Antónia deu particular relevo aos meios computacionais usados na análise dos modelos bayesianos – muitos "*likes*"!
Seguiu-se a abertura formal do congresso, onde pudemos confirmar, com satisfação, uma sala cheia. A representante da Reitoria da Universidade do Algarve, Prof.ª Lurdes Cristiano, deu-nos as boas vindas; o Presidente da Câmara Municipal de Olhão, Dr. António Pina, transmitiu-nos a sua satisfação pelo facto de poder acolher na sua cidade tão ilustre comunidade, tendo até referido um carinho especial, vindo dos tempos de estudante, pela Estatística e seus estudiosos… O representante do INE, Dr. Carlos Marcelo, aproveitou para reafirmar o orgulho na produção de matéria-prima para a investigação da comunidade estatística. A Prof.ª Eduarda Silva, na qualidade de presidente da SPE, congratulou-se pela ampla participação dos sócios em mais uma edição do congresso da SPE, marco importante no calendário das atividades de investigação em Estatística e Probabilidades, tendo também expressado o agradecimento muito especial à Comissão Organizadora Local e à Comissão Científica pelo trabalho realizado. Esse agradecimento foi ainda reforçado pelo Prof. Paulino, com algumas responsabilidades no apadrinhamento deste congresso. A Prof.ª Clara Cordeiro e a Prof.ª Conceição Ribeiro, da Comissão Organizadora, expressaram também o seu agradecimento a todas as entidades que patrocinaram o evento, tendo manifestado a sua satisfação pela enorme adesão da comunidade, patente em cerca de 160 comunicações.

Quase a terminar o dia, fomos levados a uma inesperada descoberta - a casa de um poeta, ela própria um poema…

Sonhar, é aspirar um mundo mais perfeito:
É dilatar a alma em êxtase bendito:
É deixar o que é mau, banal ou imperfeito,
Para atingir o que é suave e infinito.

O requinte da receção de boas vindas, oferecida pela Câmara Municipal de Olhão, esteve decerto à altura do sonho do poeta João Lúcio, quando projetou o chalé (que apenas ocupou por escassos meses devido a uma morte precoce, aos 38 anos).
(*in* http://www.olhao.web.pt/Textos/Biblioteca/JLucio.pdf)

Figura 1 O chalé de João Lúcio

Na quinta-feira, depois de uma manhã repleta de atividades, com sessões temáticas de Consultoria Estatística, Saúde, Estatística Bayesiana, Extremos, Séries Temporais e Modelos Longitudinais, e ainda, a apresentação do primeiro conjunto de *posters*, houve a entrega do Prémio SPE 2015 à Ana Borges, da Universidade do Minho.

Seguiu-se um passeio de barco pela Ria Formosa, pautado por duas surpresas muito agradáveis: a visita à ilha de Armona, uma ilha com cerca de 9Km de comprimento, e à ilha do Farol (construído em 1851), com uma população residente de pescadores. Uma volta pelas ruas labirínticas destes povoados deu-nos a conhecer uma outra realidade urbanística, diferente daquela a que estamos habituados.

Figura 2 O embarque

Figura 3 Aplicações da estatística

Figura 4 A Ilha do Farol

Os convidados

O painel de convidados do XXII Congresso da SPE presenteou os congressistas com oradores de excelência em diferentes e interessantes temas: na primeira sessão plenária ouvimos o Prof. Peter Muller, de origem austríaca, a trabalhar na Universidade de Texas, em Austin, que, depois de surpreender a assistência com algumas frases em português, apresentou uma comunicação sob o título "A Bayesian feature allocation model for tumor heterogeneity". O convidado seguinte veio da Universidade de Oxford, o Prof. James Taylor, que fez jus ao nome no que concerne à sua presença em palco, tendo "tocado" uma outra música – "Predicting the expected shortfall corresponding to value at risk forecasts produced by quantile models". No dia 9, foi a vez da Prof.ª Luzia Gonçalves, do Instituto de Higiene e Medicina Tropical da Universidade Nova de Lisboa, nos pôr a refletir sobre as dificuldades do trabalho de campo, com a palestra intitulada "Dados, ética e modelos na saúde tropical: constrangimentos e desafios". No último dia do congresso ouvimos ainda o Prof. Manuel Scotto, do Instituto Superior Técnico, com a comunicação intitulada "O operador *thinning* na modelação de séries temporais de valores inteiros".

A novidade

Este congresso fica marcado pela estreia da atribuição de bolsas de mérito a jovens investigadores nas áreas das Probabilidades e Estatística, por forma a incentivar a participação no congresso das camadas mais jovens.

E muito, muito mais…

O penúltimo dia do congresso foi um dia cheio de emoções! Nesse dia a manhã começou com sessões temáticas paralelas, em Ambiente e Ecologia, Educação, Ciências Sociais, Estatística Bayesiana, Séries Temporais, Demografia e Análise Multivariada, com um interregno para a segunda sessão de *posters*. À tarde as sessões de trabalhos continuaram com Análise de Correspondências, Cadeias de Markov, Modelos Lineares, Ensino Superior, Análise Multivariada, Inferência Estatística, Estudos Comparativos e Aplicações em Saúde. Ao fim da tarde desse dia assistimos a um momento muito especial e repleto de emoções.

O prémio carreira SPE foi atribuído à Prof.ª Antónia Turkman, sócia fundadora da SPE e atual presidente da Mesa da Assembleia Geral, reconhecida pela sua dedicação ao ensino e investigação da Estatística, com um percurso pleno de integridade, seriedade e generosidade. Pudemos assim testemunhar um momento cheio de simbolismo, comovente, porque não? É lícito que nos comovamos na admiração de uma vida que tanto contribuiu para o desenvolvimento do ensino e da investigação da Estatística em Portugal.

Houve ainda tempo para uma reunião da Assembleia Geral da SPE, à qual se seguiu o muito esperado jantar do Congresso.

Numa sala ampla do Real Marina Hotel decorreu um jantar bem servido, animado pelo grupo folclórico de Santo Estêvão de Tavira e que incluiu atuações de alguns congressistas.

Já é tradição nos jantares dos congressos da SPE usufruirmos de momentos de descontração e convívio, que tanto facilitam o relacionamento profissional e pessoal entre colegas.

No sábado, último dia do congresso, os trabalhos foram concluídos com sessões temáticas em Ciências Marinhas e Bioinformática, Controlo de qualidade/Extremos, Processos estocásticos/Métodos robustos, Aplicações – Economia e Séries temporais.

A sessão de encerramento foi presidida pela Presidente da SPE, Prof.ª Eduarda Silva. A Presidente agradeceu a todos a forma como decorreram os trabalhos, realçando o valor que emana da componente científica do congresso e notou o caráter intergeracional, que já é uma característica presente nos congressos da SPE; deu ainda os parabéns aos elementos da Comissão Organizadora do XXII congresso da SPE, pelo facto do encontro ter sido um sucesso.

A Prof.ª Clara Cordeiro realçou a presença de muitos estudantes, feito conseguido, nomeadamente, pela atribuição de bolsas já referida. Assinalou também a atribuição do Prémio SPE à Ana Borges, da Universidade do Minho e do Prémio Carreira, à Prof.ª Antónia. Agradeceu a oportunidade gerada com o desafio lançado pelo Prof. Paulino e pela Direção da SPE, para a organização deste congresso, e reafirmou os agradecimentos a todas as entidades que prestaram apoio institucional, à gerência do hotel, à Câmara Municipal de Olhão, ao INE e a todos os restantes patrocinadores.

Esperamos ter sido fiéis nesta descrição resumida do XXII Congresso da SPE. Parabéns à organização e votos de bom trabalho de preparação para o próximo…

Anabela Rocha
M. Cristina Miranda

XXII Congresso SPE

• Explorística vence mais um Prémio

Explorística ganhou o **Prémio ISLP`s Best Cooperative award 2015**.

Como é divulgado na página do IASE: "**Exploristica**, the result of cooperation between Sociedade Portuguesa de Estatística (SPE), Instituto Nacional de Estatistica (INE), and Ciência Viva in Portugal". *http://iase-web.org/islp/Competitions.php?p=Best_Cooperative_Project_2015.*

Neste sítio pode-se conferir mais informação sobre o referido Prémio, vencido entre 4 concorrentes.

O Prémio foi recebido por Carlos Braumnn, em representação da Presidente da SPE, na Cerimónia de entrega de Prémios do World Statistics Congress, ISI2015, realizado no passado mês de agosto no Brasil.

O mérito deste prémio deve-se, sem dúvida, aos esforços e trabalho incansável do nosso colega Pedro Campos (o coordenador) e da sua equipa.

No Boletim primavera 2015, p. 14-15, foi apresentada uma breve retrospetiva onde se descreveram "Os primeiros 720 dias" de **Explorística**, criada em 2011 e com um património que já inclui diversos Prémios.

FR

• Prémio Carreira - SPE

A Sociedade Portuguesa de Estatística, em 2013, instituiu o Prémio Carreira – SPE.

Em 2015, o Prémio Carreira – SPE foi atribuído à Prof. Maria Antónia Amaral Turkman, em reconhecimento pelas suas relevantes contribuições no desenvolvimento científico, pedagógico e de divulgação da Estatística em Portugal.

A Prof. Maria Antónia Amaral Turkman é Professora Catedrática Aposentada da Universidade de Lisboa. O Prémio Carreira – SPE 2015, em cerimónia especial, foi entregue durante o XXII Congresso SPE. Mais à frente, neste Boletim outono 2015, apresentamos um testemunho.

FR

• Prémio SPE 2015

O Prémio SPE 2015 foi atribuído a Ana Borges.

Ana Borges, é docente equiparada a Assistente na Escola Superior de Tecnologia e Gestão de Felgueiras do Instituto Politécnico do Porto.

O Júri do Prémio SPE 2015 foi presidido por Paulo Oliveira, Professor Catedrático da Universidade de Coimbra. No final deste Boletim publicamos uma notícia alargada sobre o Prémio SPE 2015.

FR

• Prémios Estatístico Júnior 2015 e Prémio aos Cursos CEF/EFA

No passado dia 3 de outubro realizou-se no Porto a Sessão de Entrega dos **Prémios Estatístico Júnior 2015** e do **Prémio aos Cursos CEF/EFA**.

No Programa foram incluídas duas palestras de Oradores Convidados:

Divertimentos com Probabilidades (e Estatística) por José Paulo Viana e *"Cientista de dados" uma profissão para o século XXI por* Irene Oliveira.

No final deste Boletim publicamos uma Notícia alargada e a Lista dos Premiados.

Após a entrega dos Prémios, o Programa integrou também a **Explorística.**

FR

• "An interview with Ivette Gomes", na Extremes

Por volta de 2012 o Miguel de Carvalho contactou-me no sentido de fazermos uma entrevista à Ivette Gomes, o que aceitei com agrado. Esta conversa teve lugar durante a preparação do Workshop EVT2013 em honra de Ivette, realizada no Vimeiro, de 8 a 11 Setembro 2013. A reunião também abraçou a celebração dos 30 anos da conferência Statistical Extremes and Applications realizada no Vimeiro em 1983, conhecida como a conferência #0 dos encontros EVA (Extreme Value Analysis), uma vez que foi pioneira nas conferências internacionais EVA. É com alegria que posso partilhar com os sócios da SPE a publicação desta entrevista no Extremes Journal da Springer, disponível no site *http://link.springer.com/article/10.1007/s10687-015-0223-3*. É incontestável que a Ivette Gomes é uma distinta cientista que imprimiu uma marca indelével e influente no domínio da Estatística de Extremos, através das suas atividades de investigação, ensino e supervisão, bem como a embaixadora da 'escola de extremos' em Portugal. Convido-vos assim a espreitar as curiosidades que são agora de domínio público e internacional nessa entrevista, online em Setembro de 2015.

Mais uma vez… Obrigada Ivette!

Isabel Fraga Alves

• 61.º Congresso Mundial da Estatística

Foi a partir do Congresso Mundial do Instituto Internacional de Estatística, *ISI 2015*, realizado no Rio de Janeiro no passado mês de agosto que, como noticiámos no anterior Boletim, p. 12, durante dois anos, a Lusofonia vai liderar o ISI – *International Statistical Institute*.
Como referido no Editorial daquele *Boletim SPE*, em conformidade, naquele congresso a nossa colega Ivette Gomes tomou posse como Vice-Presidente do ISI que, por sua vez, tem como Presidente, Pedro Nascimento Silva. É um grande orgulho para a Associação Brasileira de Estatística e para a Sociedade Portuguesa de Estatística.

O próximo Congresso Mundial de Estatísticos, na sua 61ª edição, será realizado em Marrocos e tem já informação disponível em *http://www.isi2017.org/.*

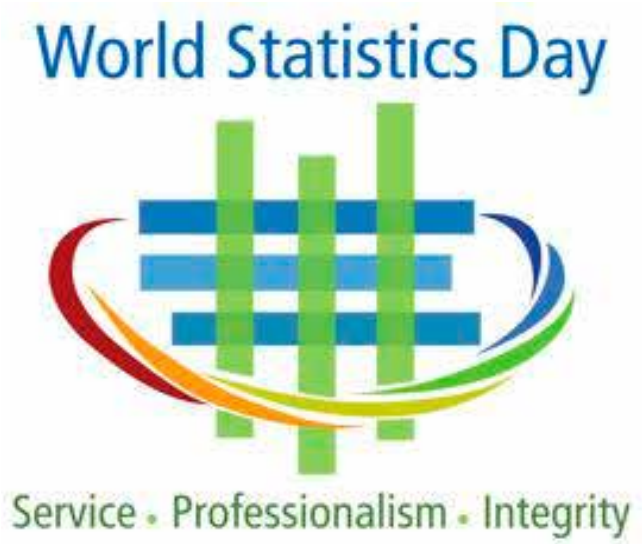
A partir de 1 de novembro já é possível propor Sessões Convidadas.

FR

• Dia Mundial da Estatística

Em *http://www.un.org/en/events/statisticsday/* as Nações Unidas anunciavam, há cinco anos:

> "Today (20 October 2010) marks the first observance of World Statistics Day, proclaimed by the United Nations General Assembly to recognize the importance of statistics in shaping our societies."
>
> *Secretary-General Ban Ki-moon*



WHY A WORLD STATISTICS DAY?

World Statistics Day (WSD) will strengthen public awareness of the important work that statisticians carry out each day. Through collecting accurate, objective and comparable data they support a wide range of national and international activities, including development efforts that improve the lives of the poor and the vulnerable.

WHAT TO EXPECT

On World Statistics Day, countries will carry out activities and events that highlight the role of official statistics and the many achievements of their national statistical systems. International and regional organizations will also hold promotional activities and events.

No passado dia 20 de outubro celebrou-se o segundo Dia Mundial da Estatística. Neste ano, o tema do WSD foi: "Melhores Dados. Melhores Vidas".

Durante o WSD a comunidade estatística mundial é solicitada a refletir e orientar-se para, através do seu trabalho, melhorar a vida da população mundial. As diversas campanhas e atividades desenvolvidas podem ser observadas em *worldstatisticsday.org*.

FR

• Bolsas para participação no XXII Congresso SPE

Pretendendo estimular o estudo e a investigação científica em Probabilidades e Estatística entre os jovens, a SPE atribuiu um número limitado de bolsas para participação no Congresso da SPE 2015, de acordo com o seguinte regulamento (do qual transcrevemos parte):

1. Os candidatos devem ser alunos de mestrado ou de doutoramento inscritos no ano lectivo 2014/2015 em alguma instituição portuguesa. São também aceites candidaturas de jovens que tenham terminado o mestrado durante o ano de 2014.
2. A bolsa é constituída pela inscrição no Congresso e uma quantia de 100 euros.
3. A candidatura consta de um resumo alargado (documento em pdf com 2 a 4 páginas) para uma comunicação oral e carta de apresentação. (…). Os candidatos devem fazer prova das condições de admissibilidade descritas em 1.
4. A decisão, da competência da Direção da SPE, será comunicada a 15 de Julho de 2015.

Com base nesta proposta da Direcção da SPE, foram atribuídas 13 Bolsas.

Testemunhos:

"A bolsa de apoio permitiu-me participar no XXII Congresso da SPE. Foram quatro dias de partilha e renovação de conhecimentos, de apresentação de resultados recentemente conquistados e projectos realizados, de reencontro e diálogo com colegas. É sempre uma experiência positiva e enriquecedora com momentos cheios de boa disposição e outros que nos emocionam".

Filipa Silva

"O XXII Congresso da SPE foi uma experiência enriquecedora e especial por se tratar do primeiro Congresso em que participo. Foi uma experiência muito rica quer do ponto de vista académico, com várias apresentações orais, plenárias e o mini curso de estatística *bayesiana*, quer do ponto de vista pessoal, uma vez que o congresso se realizou no município Olhão o que me possibilitou conhecer alguns pontos turísticos assim como a gastronomia e ainda algum contacto com a cultura algarvia pois contámos com a presença do rancho folclórico de Tavira no jantar do Congresso.

Para mim, esta participação, foi também memorável pois tratou-se da minha primeira apresentação oral. Quero aproveitar para agradecer o apoio prestado nesse sentido. Aproveito também para agradecer à direção da SPE pela bolsa que me foi atribuída e que facilitou a minha presença no congresso e, desde já, felicitar a iniciativa".

Andreia Gonçalves

"Gostaria de agradecer à direção da SPE por esta valiosa e enriquecedora oportunidade e de congratular a comissão organizadora pelo excelente trabalho desenvolvido na preparação de todas as componentes deste congresso.

O programa deste congresso era tão interessante que se tornava muito difícil escolher entre sessões simultâneas. Foi muito gratificante receber um *feedback* positivo e incentivos para continuar o trabalho que tenho desenvolvido".

Margarida Azeitona Vilela

"Foi com grande entusiasmo que participei pela primeira vez no congresso organizado pela SPE. O ambiente deste congresso, quer ao nível científico, quer ao nível das relações pessoais superou em tudo as minhas expectativas, pelo que recomendo vivamente a participação dos jovens investigadores nesta iniciativa científica. As apresentações realizadas, além de motivadas por temáticas de grande interesse, tiveram grande valor científico, o que nos permitiu aprender mais sobre esses temas. Enquanto aluno de doutoramento em matemática, a trabalhar na área das probabilidades e processos estocásticos com ênfase em finanças, gostaria de deixar a sugestão de se reforçar o número de sessões temáticas em processos estocásticos, cálculo estocástico e probabilidades, quer pelo convite de investigadores nessa área quer pela integração de mais membros com esse interesse".

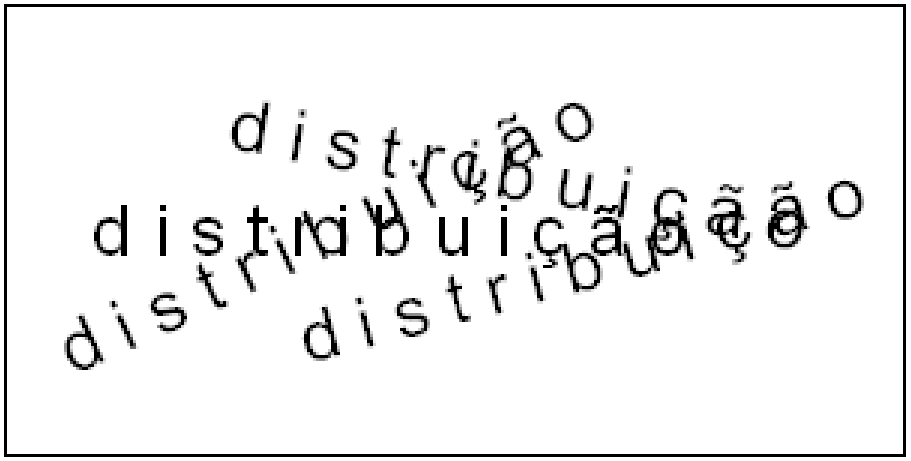
Carlos Oliveira

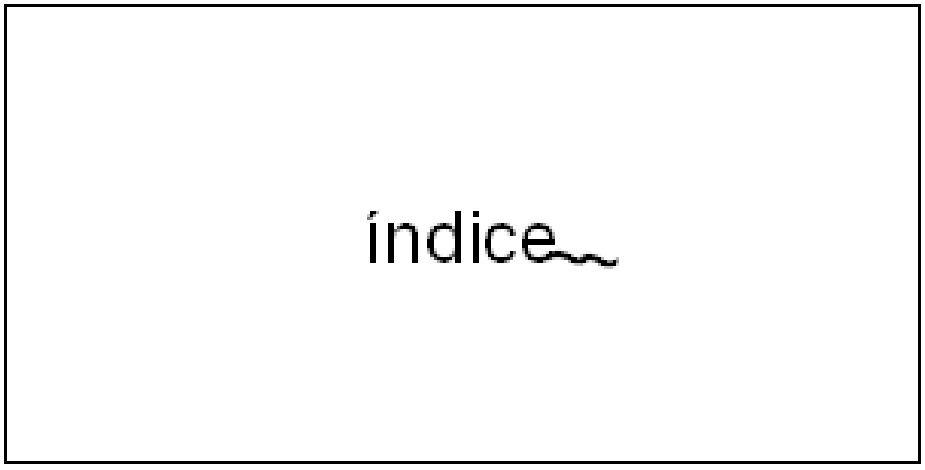
• II Encontro Galaico-Português de Biometria

No seguimento do *I Encontro Luso-Galaico de Estatística no Medio Ambiente e na Ecoloxía*, realizado em Vila Real em 2014, as Sociedades Galega para a Promoción da Estatística e da Investigación de Operacións (GGAPEIO) e Portuguesa de Estatística (SPE) estão a organizar o *II Encontro Galaico-Português de Biometria com aplicações às Ciências da Saúde, à Ecologia e às Ciências do Ambiente*. Este evento, que se irá realizar em Santiago de Compostela (Espanha), de 30 de Junho a 02 de Julho de 2016, para além de proporcionar a já excelente cooperação entre ambas as Sociedades, visa difundir os mais recentes desenvolvimentos de metodologias Estatísticas e promover a sua utilização na resolução de problemas de índole prática de diversas áreas, tais como as Ciências Naturais e do Ambiente e as Ciências da Vida e da Saúde. Este encontro contará com um vasto leque de especialistas, dos quais salientamos Geert Molenberghs, da Universidade de Hasselt (Bélgica), Daniela Cocchi, da Universidade de Bolonha (Itália), Thomas Kneib, da Universidade de Göttingen (Alemanha) e Carlos Braumann, da Universidade de Évora (Portugal). Para além de quatro sessões convidadas, um minicurso sobre Análise de Sobrevivência lecionado por Luís Machado, da Universidade do Minho (Portugal) e de uma mesa redonda, na qual serão discutidos temas de interesse comum a ambas as Sociedades, o encontro contará com numerosas contribuições orais e em forma de poster. As apresentações submetidas serão objecto de uma cuidadosa selecção por parte dos comités Científicos Português e Galego presididos por Maria Antónia Amaral Turkman, Universidade de Lisboa (Portugal) e por Carmen Mª Cadarso Suárez, da Universidade de Santiago, respetivamente. A organização do evento é da responsabilidade conjunta de ambas as Sociedades, sendo a comissão organizadora presidida por Mª Isolina Santiago Pérez, da Direção Geral de Inovação e Gestão da Saúde Pública e por Patrícia de Zea Bermudez, da Universidade de Lisboa (Portugal). Mais detalhes sobre este importante evento vão sendo disponibilizados em *http://biometria.sgapeio.es* e são acompanhadas por um vídeo promocional que poderá ser descarregado em *https://www.dropbox.com/s/ehmu6yhtvbkdwcz/VIDEO_BIOAPP2016_Gal.mp4?dl=0.*

A Comissão Organizadora

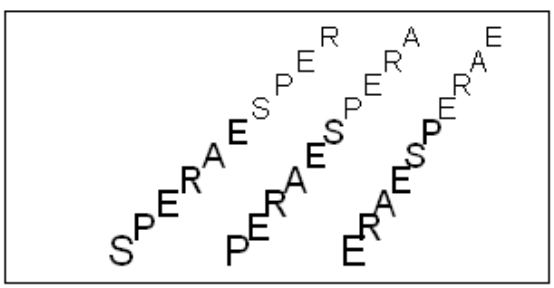
Enigmística de mefqa





No Boletim SPE primavera de 2015 (p.16):

Correlação parcial



filas de espera

Consequências da 1ª Guerra Mundial na elaboração dos livros de Probabilidade

Filipe Papança, *filipe.papanca@gmail.com*

Academia Militar

O contexto cultural exerceu uma influência decisiva no desenvolvimento da Matemática. Os problemas vividos nos campos de batalha, mormente os relacionados com o tiro de armas, motivaram o desenvolvimento da Matemática e da Estatística. Em termos didáticos tal facto motivou uma nova geração de manuais resultante de uma reflexão assente na prática dos conflitos e dos problemas com eles relacionados. Findas as hostilidades, o regresso à normalidade permitiu reunir, organizar, revelar e divulgar o novo conhecimento, entretanto, surgido.

Após a Primeira Guerra Mundial as principais obras não surgem já assinadas pelos "Matemáticos ditos influentes" (estes aparecem apenas como prefaciadores, orientadores e organizadores) mas por militares que estiveram na guerra e viveram de perto as situações ou por académicos que estiveram em contacto com eles. Por essa razão os exemplos dados não são meramente didáticos, embora essa preocupação esteja sempre presente como se pode constatar nas obras *La probabilité dans les Tirs de Guerre* (1919) de Jean Aubert de 1919 com prefácio de M. M. d'Ocane, professor da Escola Politécnica, *Applications au Tir* (1926), de J. Haag, integrando o tomo IV denominado *Applications Diverses et Conclusion* da obra de Émile Borel, *Traité du Calcul, dês Probabilités et de ses Applications* em que colaboraram igualmente matemáticos como L. Blaringhen, C. V.L. Charlier; L Deltheil, P. Dubreil, M Fréchet, H. Galbrun, F. Perrin e P. Traynard.

***Probabilité du Tir* - capitão S. Burileano**

O objetivo desta obra segundo é dito no seu prefácio é de colocar nas mãos dos oficiais de todas as armas um livro contendo o desenvolvimento completo das aplicações da teoria geral das probabilidades ao estudo experimental e à prática do tiro da espingarda e do canhão e ao mesmo tempo apresentar aos especialistas e ao grande público os princípios da dita teoria de uma forma acessível.

O primeiro e segundo capítulos contêm os princípios elementares da teoria das probabilidades, terminando ambos com resumos com o objectivo no que concerne ao grande público de adquirir sem nenhuma dificuldade as noções de cálculo das probabilidades, permitindo aos oficiais, passar rapidamente da teoria propriamente dita para o capítulo das aplicações.

No capítulo primeiro são abordados os temas: definição matemática da probabilidade, probabilidade total, probabilidade composta (definida como a probabilidade que resulta da associação de vários outros eventos, o que hoje se denomina probabilidade condicionada), probabilidade das causas, (entendendo-se como causa o conjunto de circunstâncias que assistem à produção de um evento de uma probabilidade determinada), probabilidade das provas repetidas, fórmula de Bernoulli, erro relativo, erro absoluto, erro provável e suas aplicações.

O capítulo segundo começa com uma introdução à lei de Gauss, a partir da curva de dispersão de um tiro de canhão, sendo deduzida a sua fórmula, descritas as suas características, apresentada uma tábua numérica de valores, passando-se de seguida à composição dos erros, módulo de precisão, fórmula de Fourier, dispersão em altura, caso de duas grandezas, elipses de igual probabilidade, elipse

provável, determinação da posição das elipses, acontecimentos independentes, acontecimento provável sobre um eixo qualquer, passando posteriormente à descrição dos diversos tipos de tiro e suas implicações no cálculo de trajectórias, terminando com uma série de considerações sobre a justeza e precisão.

No capítulo terceiro é apresentado um estudo experimental dos tiros de artilharia e de infantaria, aplicado à prática do tiro de diferentes armas: canhão, fuzil, obus, sendo enunciados problemas práticos, seguidos da respectiva resolução.

***Applications au Tir* - J. Haag**

O objetivo desta obra é de expor de forma sucinta e o mais completa possível, os resultados conhecidos à data sobre a probabilidade do tiro, acrescidos dos trabalhos pessoais escritos durante a guerra e que a redacção do livro obrigou o autor a terminar.

A obra começa com uma introdução à noção de dispersão aplicada à probabilidade do tiro no capítulo primeiro, sendo tecidas algumas considerações sobre o ponto médio como sinónimo de centro de dispersão, a posição provável do ponto de impacto ou de explosão; erro médio, erro médio quadrático e erro provável.

No segundo capítulo afirma que como a teoria dos erros conduziu à Lei de Gauss, é natural procurar encontrar uma solução análoga para o problema da dispersão. Este capítulo representa uma tentativa de edificação dessa teoria. O primeiro método proposto consiste numa reflexão das razões que levaram Gauss a estabelecer as Leis dos Erros, analisando essas causas e procurando estudar a teoria da dispersão a partir delas, chegar ao estabelecimento de uma teoria matemática da dispersão.

No segundo terceiro apresenta um estudo das propriedades da dispersão admitindo a Lei de Gauss. Começa inicialmente por obter as fórmulas do erro médio e do erro médio quadrático, para uma dispersão linear partindo da distribuição Gaussiana. Aplica esta lei ao estudo do tiro que explode por percussão ou seja o tiro ao alvo, cuja dispersão se efectua ao longo de uma elipse, análise probabilística essa que abrange diversas vertentes: banda rectilínea indefinida, a projecção numa direcção qualquer, o caso do paralelogramo cujos lados são direcções conjugadas, sector elíptico, ângulo ao centro, probabilidade num ângulo qualquer. Conclui que a citada Lei se aplica ao caso particular em que as elipses de probabilidade são círculos, apelidando-a de *Lei de Gauss Isotrópica*. De seguida analisa o caso do tiro que se espalha lentamente, afirmando que os pontos de igual probabilidade sobre os elipsóides homotéticos do elipsóide unitário e que apelida de elipsóides de probabilidade, afirmando, baseado em cálculos anteriores que a probabilidade de o tiro se produzir no interior de um desses elipsóides é maior do que se produzir em todo o volume equivalente. Por fim estuda o caso do prisma e do paralelepípedo, terminando o capítulo com o estudo do caso da probabilidade da projecção de uma direcção sobre um plano, concluindo que *a probabilidade das projecções dos tiros sobre uma direcção qualquer, paralelamente a um plano qualquer obedece à Lei de Gauss, com um erro unitário igual à média do segmento projectado sobre o eixo de projecção para os planos projectantes tangentes à elipsóide unitária*, chegando igualmente à conclusão de que a *projecção cilíndrica das explosões sobre um plano qualquer obedece a essa Lei com uma elipse unitária originada pela sombra projectada, quando supomos os raios luminosos paralelos à direção dos projetantes.*

O capítulo quarto trata dos tiros balísticos e das tabelas de tiro. O autor começa por estudar o efeito do alcance, afirmando que o tiro balístico comporta *n* tiros de canhão, efectuados nas condições mais uniformes possíveis, adoptando como condições teóricas certas condições médias, adoptando o princípio do tiro de canhão fictício efectuado nessas condições. Introduz nesse conceito a noção de desvio em relação à média, afirma que os erros são independentes, convergem **assimptoticamente para a Lei de Gauss**, tem um limite, uma constatação que está na base do que actualmente se denomina de **Teorema do Limite Central**, se bem que estatísticos influentes continuem a fazer questão de o apelidar simplesmente de **Teorema do Limite**. Introduz diversas noções como por exemplo, erro provável, erro médio e erro quadrático médio, incluindo nessas formulas o factor de correcção **n-1**, uma novidade em relação às outras obras anteriormente analisadas, o que segundo o autor as torna mais exactas, concluindo que o método do erro quadrático médio é ligeiramente superior ao erro médio. A lei de probabilidade do erro unitário aplicada ao método da mediana, considerado

mais rápido embora sendo menos eficaz que os anteriores e onde já surge igualmente a expressão ***assimptótica à lei de Gauss***. Efectua uma abordagem ao método das diferenças sucessivas, da autoria do capitão Bréger, aconselhável quando existe alguma perturbação, atmosférica ou outra, aumentando a zona de dispersão, modificando a vertente teórica, não obedecendo à Lei de Gauss, as regras anteriormente descritas para cálculo do erro unitário caem também. Bréger propõe então que se recorra às diferenças entre dois tiros consecutivos uma vez que neste caso o enunciado descrito anteriormente varia muito pouco, uma vez que as diferenças sucessivas obedecem à **Lei Norma**l e essa diferença é sempre constituída por grandezas independentes. A construção de tabelas de erros prováveis, os erros em direcção, as elipses de probabilidade, os erros no plano de tiro são outros dos tópicos abordados.

O capítulo quinto é dedicado ao tiro ao alvo, aquilo que foi dito em relação ao tiro que explode por percussão continua válido para esta modalidade, haverá que considerar agora um plano vertical em vez de um plano horizontal. Em particular tudo o que foi dito para a determinação do ponto médio pode ser aplicado.

Afirma que o método mais preciso seria medir as coordenadas dos pontos de impacto em relação a dois eixos rectangulares quaisquer que eles sejam e calcular a sua *média aritmética*. Se o número de pontos, for numeroso, tal operação tornar-se-á um pouco fastidiosa teremos então de nos contentar com o *método das medianas*, já anteriormente descrito. Um outro método mais expedito, segundo o autor consiste em substituir cada mediana por uma direcção equidistante dos pontos de impacto mais afastados. O ponto tomado como médio será então o centro do rectângulo mínimo contendo todos esses pontos e tendo os seus lados paralelos às duas direcções rectangulares dadas. O erro a recear será desta vez maior do que um ou outro dos métodos anteriores.

É definido o conceito de justeza no tiro como sendo a distância do ponto médio ao ponto visado. Quanto mais pequena for essa distância maior será a justeza. Em lugar do ponto médio real toma-se como ponto de referência um ponto médio fictício, cometendo-se então um certo erro de onde è fácil encontrar a lei de probabilidade. Toma por origem o ponto visado e coloca-o no centro da circunferência, de onde se assimilam os erros aos diferenciais. Admite que a lei da dispersão segue a lei de Gauss isotrópica descrita no capítulo terceiro (quando as elipses são círculos).

No caso do tiro ao alvo, a zona de dispersão surge um pouco mais circular, afirma que a experiência mostra que as elipses de probabilidade são geralmente círculos e aplica os métodos anteriormente descritos (mediana, média e média quadrática), ao tiro ao alvo, concluindo que a dispersão que lhe está associada segue a **Lei de Gauss Isotrópica** e por essa via determina o erro que lhes está associado, passando à descrição da Lei de probabilidade de razão unitária associada ao método do círculo mediano, afirma que esta é **assimptótica à Lei de Gauss**, determina o seu erro e conclui que o método do círculo mediano é muito menos bom que o método da média quadrática. Descreve o método do círculo total, um caso particular do antecedente, definindo a variável x como o raio mais pequeno contendo todos os pontos de impacto. Termina com uma avaliação da precisão do tiro, supondo a elipse um círculo.

O capitulo sexto é dedicado à regulação do Tiro, definido como a procura experimental dos elementos para os quais o ponto médio fictício coincide com um ponto médio dado à partida, apelidado de alvo ou *objectivo*. È fornecida uma regra muito simples para medir os erros: efectua-se um conjunto de tiros partindo de uma determinada posição e avaliam-se os erros algébricos (positivos ou negativos) em relação a um determinado ponto definido como objectivo. È feita uma introdução à teoria elementar da aproximação, descrevendo o método em uso pela artilharia Francesa que começa com *dois tiros de ensaio por duas peças diferente*, sendo efectuadas de seguida correcções a esse tiro e estabelecendo uma ponte com a teoria matemática rigorosa, fazendo uso da fórmula de Bayes e da Lei de Gauss. È descrito o modo de regular um conjunto de baterias (quatro segundo o exemplo), fazendo a média das médias do desempenho de cada uma das baterias, o que constitui uma aplicação do teorema do limite. O capítulo termina com uma descrição do modo de regular o tiro em função de uma ou duas direcções de referência (conjugadas ou não) e do modo de como efetuar essa correção.

O capítulo sétimo é dedicado à eficácia e ao rendimento do tiro aplicando a Lei de Poisson, e introduzindo fórmulas para avaliar esse rendimento, constituindo o capítulo oitavo uma aplicação do que foi dito no capítulo quinto ao tiro de caça.

O capítulo oitavo é dedicado à problemática relacionada com o tiro de caça, terminando a obra com uma série de notas, o que hoje poderíamos apelidar de anexos.

La probabilité dans les Tirs de Guerre de Jean Aubert com prefácio de M.M d`Ocane

No prefácio é salientado que a originalidade desta obra reside no facto de não ter sido elaborada no silêncio do gabinete, mas terá começado no fogo de acção por um jovem oficial de artilharia praticante, durante quatro anos e meio.

Em relação aos problemas que estudou de uma forma isolada, fragmentada, *Jean Aubert* resolve organizá-los sob a forma de um princípio unificador daí resultando um encadeamento lógico, tendo em vista despertar o interesse pelo estudo das questões colocadas, procurando ao mesmo tempo libertar-se de duas convenções, a primeira de que os tiros compreendem um número infinito de ocorrências e a segunda, admitir para cada um deles a existência de um ponto médio fixo e invariável, embora para efeitos estritamente didácticos, seguindo o caminho do mais simples para o mais complicado, suponha provisoriamente a existência de um ponto médio invariável. Estuda em particular os diversos métodos do tiro, a questão da precisão das determinações do erro provável de um canhão e os diversos problemas relacionados com o tiro resultante da simultaneidade das diversas peças.

No capítulo primeiro, começa por abordar a noção de *grandeza eventual* (apresentada para evitar confusões como sinónimo de experiência) que surge da noção de erro, erro esse que comporta uma medida dependente em parte do acaso, considerando este em si mesmo um valor numérico de medida, aproveitando a introdução do conceito para estabelecer logo uma relação com o tiro e sua curva de dispersão, curva de Gauss, abeirando-se do que hoje se designa por teorema do limite central.

O segundo capítulo é dedicado ao tiro em condições atmosféricas invariáveis, ou seja a temperatura e a pressão atmosférica são consideradas constantes. O vento não é considerado nulo mas somente invariável. Estas hipóteses múltiplas não suprimem as causas habituais da dispersão: velocidade inicial variável de um lado ao outro. Os obuses diferem uns dos outros pelo peso, a posição do centro de gravidade, a forma e o grau de polimento da superfície exterior. Começa por tratar o tema da dispersão em volta de um ponto médio real e aparente, procurando avaliar o grau de eficácia do tiro num alvo. Passa ao cálculo do coeficiente de eficácia relativa e à curva de frequência, utiliza diversos métodos como o do ponto médio. Procura de seguida avaliar o erro provável de um canhão, descrevendo diversos métodos como o do ponto médio aparente, o dos erros entre tiros consecutivos, terminando com uma abordagem do tiro utilizando diversas peças.

No terceiro capítulo é abordado o tema do tiro em condições atmosféricas desfavoráveis, analisando-se os efeitos das variações de temperatura, do vento, da pressão atmosférica, procurando-se medir o grau de imprecisão (erro provável) que no caso de uma peça, quer no caso de várias peças.

No quarto e último capítulo é analisada a influência de novos factores sobre o problema do tiro como sejam o factor tempo, a precisão de observação, a forma do terreno, as dimensões do objectivo, defeitos do canhão e das munições relacionando estes factores com a Lei de Gauss, passando de seguida a tratar da probabilidade resultante da combinação das diversas causas, inspirando-se no teorema de Bayes analisa as causas *a priori* e *a posteriori*, terminando com a análise de questões diversas como a maneira de arredondar a medida de um ângulo, tiro sem observação, tabelas de tiro e factores morais.

Referências Bibliográficas

Aubert, J. (1919). *La probabilité dans les tirs de guerre*. Paris: Gauthier Villars.

Burilano, S. (1911). *Probabilité du Tir*. Paris: Octave Doin et Fils, Éditeurs.

Haag, J. (1926). *Applications au Tir*. Em Émilie Borel (Ed.), *Traité du Calcul des Probabilites* (Tomo IV). Paris: Gauthier Villars.

SPE e a Comunidade

Uma visita guiada às ocupações do Laboratório de Cineantropometria e Gabinete de Estatística Aplicada da Faculdade de Desporto da Universidade do Porto.

José Maia[1], *jmaia@fade.up.pt*

com a ajuda dos colegas Rui Garganta[1], Duarte Freitas[2] e António Prista[3]
[1]CIFI[2]D, Faculdade de Desporto, Universidade do Porto
[2]Centro de Ciências Sociais, Universidade da Madeira, Funchal
[3]Faculdade de Educação Física e Desporto, Universidade Pedagógica, Maputo, Moçambique

1.**"Fotografias breves" na SPE**

Confesso que fiquei surpreendido quando recebi o honroso convite do Prof. Fernando Rosado para enviar ao Boletim da Sociedade Portuguesa de Estatística (SPE) um texto sobre Estatística no Desporto. O texto era para sair no número anterior do Boletim. Infelizmente não pude responder ao seu pedido. Combinamos que o faria neste. Muito lhe agradeço a paciência e o cuidado para não faltar à chamada. Aqui está o texto, uma espécie de "carta de navegação" de tudo quando foi realizado com a ajuda de muitas pessoas, de que destaco os Profs. Rui Garganta, Duarte Freitas e António Prista. Espero que responda ao que o Prof. Fernando Rosado tinha em mente. Contém, de certo modo, brevíssimas fotos da minha história na SPE, o que estudo e a minha paixão pela Análise de Dados Quantitativos em decorrência das pesquisas realizadas no Laboratório de Cineantropometria e Gabinete de Estatística Aplicada. Sempre com a ajuda de muitos alunos e colegas de Portugal e do estrangeiro.

Lembro-me muito bem da minha participação nos primeiros congressos da SPE, sobretudo do primeiro, realizado em Troia. Tenho bem presente o meu nervosismo por ousar navegar num mar que não era o "meu", nunca foi, nem é. Às vezes a ignorância tem destas coisas – o atrevimento. Lembro-me, também, das perguntas que me fizeram os Profs. Dinis Pestana e Galvão de Melo. Guardo na minha memória o modo sempre muito atencioso com que os membros da SPE me trataram, mesmo sabendo que eu era de "outras bandas".

Mais tarde tive a honra de trabalhar mais de perto com a Profª Rosário Oliveira e o Prof. João Branco. Levei-lhes alguns problemas na esperança que me ajudassem na sua solução. Assim foi, de modo muito eficiente, sempre rodeado de cumplicidades e amizades. Um dos problemas era precisamente sobre a análise de um teste/escala de avaliação do conhecimento declarativo do jogo de Basquetebol. Sob a orientação do Prof. João Branco, a Profª Rosário Oliveira apresentou uma solução muito elegante. Fazem-me falta as viagens que fiz ao Técnico para conversar com eles.

A minha última participação nos congressos da SPE foi exatamente no Luso. Não me lembro do ano. Assisti a praticamente todos os trabalhos. Foi nesta altura que percebi, realmente, que não estava no "meu mundo". Não obstante o elevado calibre dos trabalhos, uma parte substancial tratava de assuntos que me escapavam totalmente. Senti-me completamente perdido. Falei com o Prof. João Branco dando-lhe conta do meu desconforto e decisão de abandonar a SPE e participar nos seus congressos anuais. Não obstante todo o seu apoio às minhas presenças por levar problemas muito concretos com dados reais, nunca mais voltei. Não sei se tomei a decisão mais acertada. O que sei, tal como referi anteriormente, é que sempre fui tratado com muita deferência e compreensão por parte de todos os colegas. Tive oportunidade, também, de convidar os Profs. Dinis Pestana, João Branco,

Rosário Oliveira e Paulo Gomes para realizarem palestras na minha Faculdade sob o signo do simples e do complexo, das árvores e da floresta. Estabeleceu-se um espaço interessante de diálogo entre estatísticos, epistemólogos e investigadores do desporto para falar conjuntamente de tanta coisa de interesse mútuo.

2. **Tarefas**

O que é que tenho feito desde essa data? Permitam que refira, para começar, algo sobre mim esperando que me desculpem por "falar no singular". Leciono várias unidades curriculares na minha Faculdade: Desenvolvimento Motor (o meu grande "amor") na licenciatura, Metodologia da Investigação e Análise de Dados no mestrado, Análise Avançada de Dados Quantitativos, Análise de Dados Longitudinais e Modelação do Desempenho Motor no programa doutoral.

Por influência de alguns dos meus mentores, a Profª Corália Vicente (Universidade do Porto), os Profs Gaston Beunen (Universidade Católica de Lovaina), Robert Malina (Universidade do Texas) e Claude Bouchard (Universidade da Louisiana), sempre tive esse "bichinho" da análise de dados quantitativos. Muito devo, sobretudo, ao Prof. Gaston Beunen que já não está entre os "vivos", mas não duvido do seu sorriso a tudo quanto foi conseguido mesmo depois de ter partido. Este interesse decorre, naturalmente, da exigência em lidar com os meus próprios dados e dos alunos de mestrado e doutoramento que trabalham comigo no vasto território do Desenvolvimento Motor. Claro que este esforço suplementar obrigou-me a fazer formação específica, sempre fora do país, durante vários anos. Mais uma oportunidade para trazer esses formadores/investigadores para projetos em que estou envolvido.

De que ocupo para além da lecionação? As áreas de estudo no Laboratório que dirijo são as seguintes: (1) Desenvolvimento Motor, muito concretamente a pesquisa interativa sobre o crescimento físico, a maturação biológica, o desempenho desportivo-motor e a multiplicidade de contextos em que ocorrem. Claro que exige informação longitudinal, longitudinal-mista ou de natureza transversal. (2) Desempenho motor coordenativo e neuro-motor em crianças e jovens, a que se associa um leque de variáveis moderadas e mediadoras, de que destaco informação sobre a história gestacional, peso à nascença, e do crescimento da criança nos dois primeiros anos de vida. Trata-se de juntar, num mesmo "pacote", matérias do Desenvolvimento Motor com as da Plasticidade Fenotípica. (3) Auxologia e desempenho desportivo em jovens atletas, sobretudo no domínio da seleção e da resposta ao treino. (4) Auxologia, desempenho motor, atividade física, maturação biológica e indicadores de risco cardiometabólico em contextos culturais distintos – Portugal, Brasil, Moçambique e Peru. (5) Epidemiologia da atividade física com dados populacionais e Epidemiologia Genética com dados de gémeos e de famílias nucleares, mais concretamente a sua atividade física, composição corporal, fatores de risco cardiometabólico, nutrição, fatores do ambiente físico e construído; nestas áreas a informação que dispomos é de natureza transversal e longitudinal.

3. **Projetos/estudos e análises**

De seguida passo a referir, muito brevemente, alguns dos "produtos" resultantes de análises dos dados com alguma complexidade provenientes de projetos financiados por Câmaras Municipais, Fundação para a Ciência e Tecnologia, Banco Mundial, agências de financiamentos de outros países e financiadores privados. Nos pontos seguintes vou tentar seguir a sugestão, extraordinariamente bem conseguida por Hair et al (2010) no seu tratado sobre *Multivariate Data Analysis*, de escrever sobre análise de dados sem uma única fórmula. Aliás, estão sempre muito bem apresentadas na literatura da especialidade. Não acrescentaria nada de novo se as escrevesse aqui. Vamos então aos pontos que considero mais relevantes nesta minha "carta":

1. *Estudos auxológicos sobre diferentes indicadores do crescimento físico (altura, peso, perímetro da cintura, índice de massa corporal e percentagem de gordura corporal) e da maturação biológica.*

   Uma das grandes preocupações de Auxologistas quando lidam com informação antropométrica, concretamente a altura, peso, índice de massa corporal ou perímetro da cintura,

é tentar sumariar a vasta massa de dados (na casa dos milhares) de modo "simples" e muito elegante do ponto de vista gráfico. O exemplo mais evidente é o da construção de *cartas percentílicas* do crescimento físico humano de enorme relevância em Pediatria, Saúde Pública, Educação, Nutrição e Ciências do Desporto a que associamos informação sobre o estado da maturação biológica (Freitas et al., 2004) e previsão da estatura final (Beunen et al., 2011). Realizamos estudos na Região Autónoma da Madeira (Freitas et al., 2002), dos Açores (Maia et al., 2007a; 2007b), em Vouzela (Chaves et al., 2015), e em duas escolas de S. Tirso (Maia et al., 2012). A nossa ferramenta de análise foi o modelo LMS de Cole e Green (1992). O mais importante de toda esta "aventura" foi construir cartas de referência locais para evitar termos de situar o crescimento, qualquer que seja o indicador utilizado, das crianças e jovens Portugueses com as referências americanas. Uma outra vantagem foi a de termos comparações mais precisas para ilustrar "onde se situam os Portugueses" relativamente a outras populações Europeias e Americanas. Adicionalmente, construímos um método de previsão da estatura final de relevância no contexto desportivo, sobretudo em termos de seleção em desportos onde o valor da altura é muito importante, bem como mostramos a alteração "secular" do crescimento e indicadores ósseos da maturação biológica de crianças e jovens Madeirenses (Freitas et al., 2012). De salientar que a aventura da construção de *cartas percentílicas* também foi conduzida no Brasil (Silva et al., 2012) e no Peru (Bustamante et al., 2015) por alunos de doutoramento.

2. *Estudos sobre o desempenho motor de natureza normativa e diferencial.*

No desporto de alto rendimento, a categoria central da sua expressão é sem sombra de dúvida a da "excelência" do desempenho. Claro que esta categoria percorre toda a atividade humana. É muito realçada no setor educativo, principalmente na disciplina de Educação Física e no desporto infanto-juvenil. Provavelmente uma das perguntas mais recorrentes de um qualquer aluno ou jovem atleta ao ser avaliado numa qualquer prova de natureza motora (habilidades e/ou aptidões) é a seguinte: afinal, o que é que o número que expressa o seu desempenho quer dizer? Dispomos de duas abordagens, designadas de normativa e criterial, para responder à questão. Uma e outra encerram uma história interessante do ponto de vista do seu desenvolvimento conceptual e matemático-estatístico.
Ora a interpretação do desempenho de crianças e jovens no contexto escolar e/ou desportivo obriga, também, à construção de *cartas percentílicas* uma vez mais com o modelo LMS de Cole e Green (1992). Realizamos esta tarefa em Portugal (Maia et al., 2007b; Chaves et al., 2014), no Brasil (Silva et al., 2012) e no Perú (Bustamante et al., 2012). Adicionalmente tentamos responder a, pelo menos, cinco problemas fulcrais na interpretação do desempenho motor: (1) o da consideração do fator tamanho (do inglês *scaling*, ou *allometry*) dado crianças e jovens da mesma idade cronológica e sexo terem dimensões corporais completamente distintas que condicionam a sua performance motora qualquer que seja o local geográfico que consideremos, por exemplo no Peru (Bustamante et al., 2015) e Portugal/Moçambique (dos Santos et al., 2015); (2) o da estabilidade (do inglês *tracking*) do desempenho motor ao longo do tempo quando a informação é de natureza longitudinal (da Silva et al., 2013); (3) o da associação do desempenho motor a aspetos socio-económicos e da maturação biológica (Freitas et al., 2007), e o da tendência secular (dos Santos et al., 2015) cujo impacto em termos de saúde e estilo de vida das crianças e jovens é relevante; (4) Se é verdade que "para quem tem um martelo tudo que tem diante de si são pregos" então, a visão que refere a estrutura hierárquica ou organizacional de praticamente toda a informação exige a consideração de modelos de análise adequados (em inglês *mixed-models*, *hierarchical linear models*, ou *multilevel models*). Podemos, mas não devemos, investigar o desempenho motor das crianças independentemente do seu contexto. Daqui que tenhamos recorrido a este tipo de modelos para interpretar o desempenho motor de milhares de crianças de largas dezenas de escolas do 1º ciclo do ensino básico da Região Autónoma dos Açores (Maia et al., 2002), de Amarante (de Sousa et al., 2005), de Vouzela (Chaves et al., 2012) e da Maia (Maia et al., 2009); (5) Não duvidamos que as crianças e os jovens são mais diferentes que iguais no seu crescimento e desenvolvimento (motor). Ora este facto coloca sérios problemas no desenho e implementação dos planos anuais de Educação Física bem como no planeamento do treino. O problema central

é o da determinação do nível de Prontidão Motora de crianças e jovens. Esta matéria foi explorada, sobretudo na Educação Física escolar, a partir da Análise da Função Discriminante e da Regressão Quantil (Maia et al., 2002; de Sousa et al., 2005; Malafaya, 2013).

3. *Pesquisa de natureza longitudinal sobre o desempenho coordenativo*

A essência do Desenvolvimento Motor é, precisamente, o da investigação sobre a mudança que ocorre na criança e no jovem numa variedade de domínios desde o físico ao motor, sempre em contextos muito precisos que são a escola, o clube ou o local de residência (ambiente natural e/ou construído). A análise de dados longitudinais coloca sérios desafios ao investigador, sobretudo quando percorre as etapas sugeridas por Baltes e Nesselroade (1979): (i) identificação da mudança intra-individual; (ii) identificação das diferenças (ou similaridades) entre indivíduos na mudança intra-individual; (iii) análise das inter-relações na mudança intra-individual; (iv) análise das "causas" da mudança intra-individual; (v) análise das "causas" das diferenças entre indivíduos na mudança intra-individual. Este fascínio levou a que fossem escritos alguns textos de "entrada" nestas matérias para investigadores do universo das Ciências do Desporto de que destaco os que se referem a modelos de estruturas de covariância (Maia et al., 2009), *mixed-models* (Maia et al., 2007; 2010), bem como ao uso de procedimentos variados para estudar o *tracking* do desempenho motor (Maia et al., 2001; 2002; 2003).

Duas das pesquisas mais emblemáticas em que me envolvi até hoje foram realizadas nas Regiões Autónomas dos Açores e da Madeira. A primeira teve por título "Crescimento somático, maturação biológica, composição corporal, coordenação motora, aptidão física referenciada pela saúde, actividade física e motivação para a prática desportiva. O estudo longitudinal-misto da Região Autónoma dos Açores" (Maia et al., 2003; 2007a). Mil crianças e jovens distribuídos por quatro coortes com sobreposição foram seguidos durante 4-5 anos num vasto leque de variáveis. A estabilidade da coordenação motora foi estudada com os procedimentos desenvolvidos por Foulkes e Davies (1981), a mudança com base em *mixed-models* (de Deus et al., 2010) e o mesmo aconteceu com a sua associação aos níveis de atividade física (Lopes et al., 2011; de Souza et al., 2014) e índice de massa corporal (Martins et al., 2010; Lopes et al., 2012).

A segunda pesquisa realizada na RAM com centenas de crianças teve por título "*Madeira Child Growth Longitudinal Study*", sendo a continuidade de uma outra de impacto na região – "Crescer com Saúde na Região Autónoma da Madeira" (Maia et al., 2013). Os principais resultados sobre a complexidade da mudança coordenativa, os seus preditores, bem como a estabilidade do desempenho de crianças estão divulgados em várias publicações de que destaco somente três (Antunes et al., 2015a; 2015b; Freitas et al., 2015).

Um dos desafios que temos em mãos é um estudo longitudinal-misto em curso com seis coortes de crianças do ensino básico da região de Vouzela. Percorre as complexidades do crescimento físico, do desempenho motor e coordenativo, bem como do desempenho cognitivo e escolar a que se associa um vasto "lençol" de preditores (fixos e dinâmicos) situados ao nível da criança, da escola, da família e do ambiente físico e construído. Uma "verdadeira aventura" que estimulará a nossa imaginação para extrairmos da enorme massa de dados as suas maiores preciosidades. A Ana Reys tem diante de si esta grande tarefa que culminará na sua tese de doutoramento.

4. *Pesquisa de Epidemiologia Genética com dados gemelares e familiares*

Por influência do meu principal mentor, o Prof. Gaston Beunen, iniciei na companhia de vários colegas e alunos de mestrado e doutoramento, um conjunto de estudos em gémeos de várias regiões do país. Claro que o maior desafio foi encontrar os gémeos, apesar de estar estimado que aproximadamente 1% da população é gemelar. Realizamos vários encontros de gémeos em várias localidades do país, tentando "misturar" aspetos da cultura de cada região, a prática desportiva e a avaliação dos gémeos. Esta aventura começou num levantamento dos gémeos e suas famílias da Região Autónoma dos Açores originando o primeiro volume escrito em Portugal sobre o assunto – "actividade física e aptidão física associada à saúde – um estudo de epidemiologia genética em gémeos e suas famílias realizado no arquipélago dos Açores" (Maia

et al., 2001), onde são apresentados, também, aspetos elementares da metodologia de análise de dados gemelares. Com financiamento da FCT sobre um projeto mais vasto, nasceram seminários internacionais, palestras, teses de mestrado e doutoramento, artigos e livros. Destaco três livros que escrevemos e que resultaram, também, de parcerias com várias Autarquias: "avaliação multimodal da actividade física – um estudo exploratório em gémeos monozigóticos e dizigóticos" (Oliveira e Maia, 2002), "o código relacional na actividade física e aptidão física associada à saúde – efeitos genéticos e ambientais" (Fernandes et al., 2006), "fatores genéticos e ambientais nos níveis e padrões de actividade física – um estudo em gémeos" (Sapage et al.,2007).

A extensão destas preocupações "rumou" para a Região Autónoma da Madeira, e com base num novo financiamento da FCT nasceu o projeto GEAFAS - "*Genetic and environmental influences on physical activity, fitness and health: the Madeira family study*" dirigido pelo meu colega de longa data – o Prof. Duarte Freitas. Foram localizadas todos os gémeos da região; convidamos toda a sua família (até à 3ª geração) e daqui resultaram vários seminários internacionais, palestras, artigos e teses de mestrado.

Alguns dos principais textos que escrevemos com informação gemelar tratam dos seguintes temas: (1) metodologias de análise de dados gemelares (Maia et al., 2001; 2011); (2) níveis e padrões da atividade física (Oliveira e Maia, 2002; Maia et al., 2002; 2003), sobretudo com o uso da entropia aproximada (Lima et al., 2010); (3) efeitos genéticos nos níveis de aptidão física associada à saúde (Maia et al., 2003); (4) validação de um questionário para determinação da zigotia relativamente a informação de micro-satélites espalhados pelo DNA (Maia et al., 2007); (5) associação do desempenho neuro-motor de gémeos com o seu peso ao nascer (Lopes et al., 2013); (6) variabilidade no desempenho coordenativo e sua associação à gemelaridade (Chaves et al., 2012). Finalmente, um resumo atual desta vasta plêiade de estudos pode ser encontrado em Maia et al. (2012).

Também sob influência do Prof. Gaston Beunen, estendi o meu interesse dos gémeos para o estudo de famílias nucleares – pais e filha(o)s (somente duas gerações). A análise de dados familiares não é "fácil" e desta vez tratamos de trazer para o nosso lado especialistas em *Statistical Genetics/Genetic Epidemiology* dos EUA (John Blangero, Vincent Diego e Peter Katzmarzyk) e de França (David Trégouët). As variáveis que tínhamos, e temos em "carteira", são variadas: crescimento físico, composição corporal, síndrome cardiometabólica, atividade física (avaliada de forma multimodal), comportamentos nutricionais, local de residência, variada informação sobre o estatuto socio-económico das famílias, a que acrescentamos uma lista extensa de dados sobre a gestação e peso ao nascer dos filhos. Temos várias amostras de famílias que vão de 1000 casos até 12000 sujeitos, em função das variáveis.

Com a ajuda dos colegas americanos e das suas propostas de análise para este tipo de dados (Blangero e Almasy, 1997; Almasy e Blangero, 1998; Diego et al., 2003; 2007; Blangero, 2009; Blangero et al., 2013), vários alunos de mestrado e doutoramento envolveram-se no trabalho em colocar os seus esforços em "montras apetecíveis". Destaco o Rogério Fermino (Fermino et al., 2008; 2009), a Michele Souza e (Souza et al., 2011), o Daniel Santos (dos Santos et al., 2012; 2013a; 2013b; 2014), a Raquel Chaves (2013), e a Thayse Gomes (2013). Acrescento a proposta de David Trégouët e Laurence Tiret (2000) para usar a teoria das *estimating equations* para lidar com correlações familiares (Maia et al., 2013; Santos et al., 2013).

Toda esta "aventura" com dados familiares foi conseguida porque obtivemos financiamento e apoio logístico da FCT, Regiões Autónomas da Madeira e Açores, Câmaras Municipais, Juntas de Freguesia e Agrupamentos de Escolas de várias regiões do país. Deixamos livros de divulgação pública deste vasto manancial de dados em vários locais: "actividade física e componentes da síndrome metabólica – um estudo em famílias nucleares, na RAA em 2005; "combata e síndrome metabólica – cuide da sua família e faça actividade física, na RAA em 2007; "Camacha saudável – um estudo em famílias de Santa Cruz", na RAM em 2008; "cada vez mais ativo (II): uma história com muitas voltas" em Vouzela em 2012; "jogos de luz no S. Tirso COM vida – uma história com 3 anos" em S. Tirso em 2012.

Finalmente para referir, neste ponto, que nos aguardam os dados de famílias nucleares Peruanas (cerca de 234, com um total de 1200 sujeitos), bem como 1100 pares de irmãos

provenientes de três regiões: nível do mar, selva amazónica e altitude (4000 m). Os apaixonados por interações entre fatores genéticos e ambientais têm aqui um belíssimo presente.

5. *Investigação no domínio da Epidemiologia da atividade física e aptidão física*

Sob o impulso de um dos meus mentores mais recentes, o Prof. Peter Katzmarzyk do *Pennington Biomedical Research Center* dos EUA, envolvi-me num projeto à escala global designado de ISCOLE (Katzmarzyk et al., 2013) – *International Study of Children Obesity, Lifestyle and the Environment*. Compreende crianças de 10 anos de idade de 12 países de cinco continentes, colhendo uma vasta teia de variáveis organizadas por vários domínios: crianças, escola, família e ambiente construído. Convocam, desde já, o uso extenso dos *mixed-models* face à estrutura organizacional da informação, pelo menos em três níveis: crianças, escolas e países. Os dados Portugueses permitiram desde já a realização de duas teses de mestrado e uma de doutoramento. As análises estatísticas necessitaram de várias abordagens de que destaco: *mixed-models* univariados e multivariados (Gomes et al., 2014a; 2014b; 2015b), regressão de *Poisson* (Borges et al., 2015) e classes latentes (Pereira et al., 2015). Com a ajuda do Prof. Donald Hedeker, um bio-estatístico americano da Universidade de Chicago, usamos uma técnica que desenvolveu designada por *mixed-effects location scale model* cuja visibilidade é bem patente numa área de estudo designada por *Ecological Momentary Assessment*, *Intensive Longitudinal Methods* ou *Daily Life Study*. A sua elegância e flexibilidade estão bem expressas nalguns dos trabalhos publicados pelo Prof. Hedeker e sua equipa (Hedeker et al., 2008; 2009; 2012; 2013; Li e Hedeker, 2012; Pugach et al., 2014; Kapur et al., 2015). Um dos primeiros artigos que escrevemos, conjuntamente, tinha precisamente por título: *Why are children different in their daily sedentariness? An approach based on the mixed-effects location scale model* (Gomes et al., 2015).

Partindo do modelo ecológico frequentemente utilizado em Epidemiologia da atividade física, bem como do modelo de Bouchard e Shephard para interpretar as relações que existem entre o sujeito, as suas aptidões e comportamentos, estilos de vida e fatores ambientais bem como o seu património genético, desenvolvemos o "*the Oporto mixed-longitudinal growth, health and performance study*" que tem uma amostra de cerca de 7000 crianças e jovens dos 10 aos 18 anos de idade seguidos durante três anos consecutivos e avaliados num vasto leque de variáveis: crescimento físico, composição corporal, atividade física, aptidão física, indicadores cardiometabólicos, comportamentos alimentares, informação parental diversa, e hábitos de sono. O desafio de realizar esta grande tarefa coube à Michele Souza que a colocou nalgumas "vitrines" importantes da nossa área (Souza et al., 2014a; 2014b; 2015a; 2015b). O nosso grande aliado foi o *mixed-model* e a sua extraordinária versatilidade para lidar com dados com uma estrutura longitudinal-mista (i.e., com *missings-by-design*).

Finalmente, mas não em último, cabe referir os trabalhos realizados ao longo de 20 anos com o meu colega Prof. António Prista da Universidade Pedagógica de Maputo em Moçambique, sobretudo o grande projeto financiado pelo Banco Mundial – Variabilidade Biológica Humana e o seu vasto leque de variáveis de diferentes domínios: crescimento físico, maturação biológica, composição corporal, atividade física, aptidão física, nutrição, parasitologia, imunologia, malária, fatores de risco cardiovascular e genética. Com certeza que gastaria algumas páginas a descrever os métodos de análise estatística utilizados, os livros escritos, os artigos publicados e as teses de mestrado e doutoramento realizadas. Fica para uma outra oportunidade.

4. **Bibliografia**

Face à sua extensão não a incluí nesta "carta". As e os interessados podem contactar-me.

Estatística em biologia molecular: o passado, o presente e o futuro

Lisete Sousa, *lmsousa@fc.ul.pt*

Faculdade de Ciências da Universidade de Lisboa, CEAUL

Carina Silva, *carina.silva@estesl.ipl.pt*

Escola Superior de Tecnologia da Saúde de Lisboa, CEAUL

Generalidades

Vivemos na era mais mensurável da história. Na era do *petabyte* (1000 *terabytes*) o desafio não é mais o armazenamento de dados, é dar-lhes sentido.

Sendo esta a era da revolução dos dados, a respetiva análise torna-se parte integrante de várias ciências. Por exemplo, a biologia molecular deixa de ser uma ciência onde os biólogos estudam um gene de cada vez, para passar a produzir milhares (agora milhões) de medições por amostra para analisar. Além disso, ao contrário da análise do ADN, que é estática, a análise da expressão genética é dinâmica, uma vez que nos vários tecidos expressam-se genes diferentes. O geneticista John Craig Venter, sequenciava organismos isolados, mas com o aparecimento de novas tecnologias e computadores com elevada capacidade de memória, que permitem a análise de dados bastante complexos, passou a estudar ecossistemas inteiros: sequenciação dos microorganismos do oceano, desde 2003, e do ar, desde 2005 (Yooseph *et al*., 2013).

A complexidade dos dados é ainda potenciada pelas novas tecnologias que, ao surgirem, são ainda pouco exploradas, produzindo dados com mais ruído dos que as anteriores. Esta complexidade e grau de variabilidade fazem com que a estatística seja um importante e inequívoco contributo na análise. Na realidade, o papel da estatística na biologia molecular vai além de uma mera intervenção. Trata-se de um pilar indissociável desta ciência! A estatística tem vindo a conquistar o seu espaço nesta nova área, tornando-se uma componente essencial de mérito reconhecido (Durbin *et al.*, 1998; Ewens e Grant, 2001).

Um estatístico que se dedica a estudos na área da biologia molecular, tem que ter a capacidade de utilizar as mais diversas metodologias estatísticas, para além de adquirir conhecimentos biológicos e computacionais. Muitas vezes, são metodologias recentes ou até mesmo metodologias que não utiliza frequentemente e que, por isso, implicam um estudo aprofundado. Por vezes, os dados têm uma natureza tão complexa que nem sequer há métodos estatísticos adequados para proceder à sua análise. Cria-se aqui uma janela de oportunidade para avanços na investigação e na produção científica na área da estatística. O contributo da estatística tem sido relevante em problemas tão distintos como, a identificação de genes com expressão diferencial sob duas (ou mais) condições experimentais diferentes, a identificação de grupos de proteínas que se relacionam (*clusters*), a classificação de indivíduos em função do tipo de cancro, *etc.* (Silva *et al*., 2016). Ainda no campo da biologia, mas fora da especificidade molecular, a estatística tem sido utilizada para estudar problemas tão diversos como filogenia, mapas físicos, metabolómica, entre outros. Refira-se também que as cadeias de Markov escondidas são cada vez mais aplicadas em biologia molecular, como por exemplo: predição da estrutura secundária de proteínas, identificação de padrões sequenciais em sequências de ADN e de proteínas, alinhamento de sequências, *etc*. (Sousa, 2007). No entanto, isto só é possível depois de se

ultrapassarem as dificuldades de comunicação entre o biólogo e o estatístico, o que acontece geralmente porque os estatísticos possuem poucos conhecimentos em biologia molecular e os biólogos poucos conhecimentos em estatística (Bang *et al.*, 2010).
Importa referir que as conclusões inerentes aos estudos não constituem um primeiro passo que ajudará os investigadores a prosseguir com a experiência de forma mais eficaz, com menor custo e tempo. Outro facto a realçar prende-se com a extrema dificuldade existente na modelação de dados biológicos, sendo frequentemente necessário admitir pressupostos que nem sempre se verificam. A frase de Einstein "As far as the laws of mathematics refer to reality, they are not certain, and as far as they are certain, they do not refer to reality", reflete esta dificuldade.

PASSADO – Um pouco de história… recente

Os crescentes avanços em biologia molecular são consequência da obtenção massiva de dados biomédicos e biológicos. Essa explosão de dados deu-se sobretudo a partir dos anos 90, nomeadamente, com os avanços na tecnologia de sequenciação de ADN e de proteínas. O acesso a grandes quantidades de dados, impulsionou o desenvolvimento de novos programas e metodologias que permitem recolhê-los, organizá-los e analisá-los, de forma a extrair toda a informação possível.
Um dos primeiros estatísticos a trabalhar nesta área de aplicação foi Terry Speed, presidente da Divisão de Bioinformática, do *Walter and Eliza Hall Institute of Medical Research*, em Melbourne - Austrália, que viu o seu vasto trabalho na área da bioinformática reconhecido em 2013 com a atribuição do prémio australiano *Prime Minister's Prizes for Science*. Entre outros, o seu contributo em vários julgamentos (como o de O. J. Simpson, por exemplo) enquanto estatístico especialista na análise de dados forenses, faz dele um dos estatísticos mais conceituados na área das aplicações à biologia molecular e da bioinformática (Gill, 2013). Tal como Terry Speed, também Simon Tavaré é um estatístico com provas dadas no mundo da bioinformática. Professor na Universidade do Sul da Califórnia e na Universidade de Cambridge, para além de dirigir o *Cancer Research UK Cambridge Institute*, Simon Tavaré tem liderado vários grupos, projetos e doutoramentos em estatística/bioinformática. Sandrine Dudoit, professora na Universidade da Califórnia, foi aluna de doutoramento de Terry Speed em 1999 e continua desde então a fazer investigação na área da bioinformática. Autora de vários livros e publicações em revistas científicas da especialidade, é um nome que dispensa apresentação neste meio.
Os nomes referidos, são apenas alguns, entre muitos. A comunidade estatística portuguesa teve o privilégio de ouvir Terry Speed e Simon Tavaré no WSGP2005 – *Workshop on Statistics in Genomics and Proteomics*, o primeiro *workshop* internacional do género realizado em Portugal, organizado pelo Centro de Estatística e Aplicações da Universidade de Lisboa. Simon Tavaré esteve ainda presente no *Follow-up Meeting* do WSGP2005, no ano de 2007, em Coimbra. Estes eventos foram antecedidos por um ciclo de seminários sobre Estatística em Genética, organizado pelo Departamento de Estatística e Investigação Operacional da FCUL, os quais foram apresentados nas Universidades de Aveiro, Évora e Lisboa, em 2002. Desde então, vários encontros internacionais têm acontecido um pouco por todo o país.
Estes investigadores, bem como todos aqueles que iniciaram esta aventura de trabalhar quantidades "astronómicas" de dados, depararam-se com os *microarrays*, que consistem numa técnica experimental da biologia molecular que procura medir os níveis de expressão de genes, em larga escala. Esta técnica inovadora impulsionou fortemente a geração de dados de elevadas dimensões e proporcionou o desenvolvimento de muitos métodos estatísticos e algoritmos computacionais. Como em outras áreas, a estatística deve intervir logo na fase do planeamento experimental, que pode ser bastante complexo. Quando a expressão genética é convertida em números, é necessário realizar o pré-processamento dos dados. O pré-processamento consiste (1) na limpeza dos dados (remoção dos valores que são consequência de anomalias inerentes à técnica experimental), (2) na transformação dos dados (com o obetivo principal de uniformizar a variabilidade dos níveis de intensidade e simetrizar as respetivas distribuições) e (3) na normalização entre *microarrays* e, em situações experimentais particulares, dentro do *microarray* (corrigir flutuações decorrentes de fatores técnicos e reter, tanto quanto possível, variações biológicas decorrentes de fatores biológicos). Ainda no seguimento do ponto (1), a remoção de valores conduz à presença de valores omissos e, consequentemente, vários métodos de imputação têm sido propostos (Celton *et al.,* 2010). Outra matéria que tem sido alvo de

investigação por parte dos estatísticos diz respeito à problemática dos testes múltiplos. A tecnologia de *microarrays* permite a análise de milhares de genes em simultâneo e, tipicamente, para cada gene é aplicado um teste de hipóteses. Estes são alguns dos exemplos onde se registou um franco avanço na investigação impulsionado pelos *microarrays*.

PRESENTE – A atualidade

Com a constante evolução tecnológica, os estatísticos deparam-se agora com um novo desafio: os dados de NGS (*Next Generation Sequencing* – sequenciação de nova geração). Esta inovadora tecnologia de sequenciação oferece uma nova oportunidade para sequenciar todo o genoma, permitindo explorar o papel das variantes genéticas raras e mutações associadas a determinadas doenças. Os estudos relacionados com a sequenciação de vários genes, sequenciação do exoma (fração do genoma que codifica os genes) e sequenciação do genoma, continuam a ser realizados embora de forma mais aprofundada. Dados de sequenciação epigenética (Bird, 2007) também têm sido disponibilizados para estudar a regulação e funcionamento dos genes. Várias bases de dados genómicos, construídas no âmbito de projetos como *HapMap* (*http://hapmao.ncbi.nlm.nih.gov*) e *1000 Genomes* (*http://www.1000genomes.org*), têm sido disponibilizadas com livre acesso, para que deste modo se promova a respetiva análise. Os dados produzidos por esta técnica de sequenciação revolucionária, colocam à disposição dos estatísticos um vasto leque de perguntas e problemas extremamente complexos. Esta realidade, abre caminho para o desenvolvimento de novas metodologias e, assim, os estatísticos poderão dar um contributo substancial em avanços científicos e tecnológicos importantes para a sociedade, uma vez que grande parte destes estudos têm implicações na área da saúde.

A necessidade de analisar esta quantidade massiva de dados de uma forma relativamente simpática e intuitiva, conduziu ao aparecimento de um projeto bem conhecido por parte de quem trabalha na área da análise de dados genéticos: o Bioconductor (*http://www.bioconductor.org*). Trata-se de uma plataforma *on-line* de livre acesso que tem por base a linguagem de programação estatística R. Sem o apoio dos meios informáticos, dificilmente se assistiria a este desenrolar de novos métodos estatísticos. Esta plataforma em particular, permitiu a disseminação da estatística e o impulsionamento na carreira de vários jovens estatísticos.

Contrariamente aos dados de *microarrays*, os dados de NGS consistem em contagens. Desta forma, o desafio associado a estes dados experimentais é muito maior, não só pela complexidade das experiências (difíceis de entender para um estatístico) mas pelas limitações inerentes aos dados discretos e amostras pequenas. Ao fim de poucos anos, após os desafios propostos pelos *microarrays*, surge novamente a necessidade de se desenvolver novos métodos estatísticos. Os biólogos não conseguem acompanhar a rapidez com que os métodos surgem, havendo tendência para usarem os métodos habituais, ou seja, os mais simples, que na maioria das vezes não são os mais adequados. Felizmente, o paradigma tem vindo a mudar e é já muito frequente haver coautores estatísticos que, com o seu conhecimento científico, imprimem uma mais-valia na análise dos dados.

FUTURO – Ir mais além…

Um problema emergente na biologia molecular consiste em integrar diferentes tipos de bases de dados como: SNP (*Single Nucleotide Polymorphism* - polimorfismo de nucleótidos simples), expressão genética, metilação do ADN, integração de informação genómica proveniente de *pathway analysis* e *network analysis*. Neste sentido, torna-se fundamental treinar a nova geração de cientistas desta era "ómica", com uma abordagem integradora de modo a dar respostas aos desafios. O treino tradicional em bioestatística não preenche totalmente as necessidades. Assim, é importante preparar bioestatísticos com várias competências. Espera-se que estes novos cientistas tenham (1) elevados conhecimentos estatísticos e computacionais de modo a desenvolverem métodos específicos para bases de dados de elevada, (2) conhecimentos biológicos suficientes (3) capacidade de trabalhar num ambiente de pesquisa multidisciplinar (4) boa capacidade de comunicação para que deste modo ultrapassem as fronteiras nos fóruns de discussão biomédicos.

Tantos requisitos demonstram que o treino integrado é essencial. Para preparar esta nova geração, é necessário desenvolver um currículo interdisciplinar de largo espetro. A oferta em Portugal não é

grande e a que existe nem sempre é conhecida de todos. No programa de treino em Bioinformática (GTPB – *Gulbenkian Training Programme in Bioinformatics*) do Instituto Gulbenkian de Ciência (Fernandes, 2010), procura-se fornecer aos participantes as competências necessárias para trabalharem com estatística, biologia, informática, *etc.*, que, todas juntas, constituem a bioinformática (Arhipova, 2006). Ao nível da formação universitária pós-graduada, existem dois mestrados na área da bioinformática, um na Universidade do Minho e outro na Faculdade de Ciências da Universidade de Lisboa (FCUL). Existe ainda na FCUL, o mestrado em Bioestatística, que oferece a possibilidade de uma formação mais vocacionada para aplicações na área da biologia molecular para os alunos que assim desejarem. Os investigadores que já têm alguma experiência no campo das aplicações à biologia molecular, também sentem necessidade de frequentar cursos de atualização.

Os estatísticos portugueses que fazem investigação em genética ou em biologia molecular encontram-se dispersos, não só em Portugal, mas também no estrangeiro. Assim, seria interessante caracterizar e identificar os colegas que investigam nesta área. A vantagem de identificar esta comunidade, para além da troca de experiências e conhecimentos entre colegas, seria darmo-nos a conhecer à comunidade de investigadores nas áreas da biologia molecular e genética. Nesse sentido, desenvolvemos um pequeno inquérito disponível em *https://qtrial2014.az1.qualtrics.com/SE/?SID=SV_8FUl9djxncGMKbj,* de modo a tentar caracterizar a comunidade de bioestatísticos portugueses que trabalham nesta área específica de aplicação. Com este inquérito pretende-se criar uma rede para promover um canal de partilha e de divulgação entre investigadores nesta área.

Acreditamos que a alteração no arquétipo dos dados só agora começou. Bases de dados que estão a ser construídas, ou mesmo as já existentes, podem esconder realidades que merecem, e devem, ser descobertas. Estas descobertas irão, certamente, levar a um maior conhecimento científico do nosso mundo. Os estatísticos devem sentir-se empolgados e preparados para desempenhar um importante papel no renascimento científico impulsionado pela revolução na medição.

Referências

Arhipova, I. (2006). The role of statistical methods in Computer Science and Bioinformatics. Em *Atas da XII International Conference on Teaching Statistics (ICOTS-7)*, Salvador - Brasil.

Bang, H., Zhou, X.K., van Epps, H.L., Mazumdar, M. (2010). *Statistical Methods in Molecular Biology*. Springer.

Bird, A. (2007). Perceptions of epigenetics. *Nature* 447: 396-398.

Celton, M., Malpertuy, A., Lelendais, G. e Brevern, A.G. (2010). Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. *BMC Genomics*, 11:15.

Durbin, R., Eddy, S.R., Krogh, A. e Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Ewens, W.J. e Grant, G.R. (2001). *Statistical Methods in Bioinformatics: An Introduction*. John Wiley & Sons Ltd., England.

Fernandes, P. (2010). The GTPB training programme in Portugal. *Briefings in Bioinformatics*, 11(6):626-34.

Silva, C., Freitas, A., Roque, S. e Sousa, L. (2016). Arrow Plot and Correspondence Analysis Maps for Visualizing the Effects of Background Correction and Normalization Methods on Microarray Data. Em *Pattern Recognition in Computational Molecular Biology: Technologies and Approaches* (M. Elloumi, C.S. Iliopoulos, J.T.L. Wang e A.Y. Zomaya, Eds), John Wiley & Sons. *In Press*.

Sousa, L. (2007). A estatística ao encontro da genética. Em *Estatística e Qualidade na Saúde: Problemas e Temáticas*, (Cunha,G. e Varanda,J., Eds.), EQS2006, p. 174–180.

Gill, A. (2013). *Statistician Professor Terry Speed wins 2013 PM's Prize for Science.* Walter and Eliza Hall Institute of Medical Research, *News*, 15 de outubro de 2013. Disponível em: *http://www.wehi.edu.au/news/statistician-professor-terry-speed-wins-2013-pm's-prize-science*

Yooseph, S., Andrews-Pfannkoch, C., Tenney, A., McQuaid, J., Williamson, S., Thiagarajan, M., Brami1, D., Zeigler-Allen, L., Hoffman, J., Goll, J.B., Fadrosh, D., Glass, J., Adams, M.D., Friedman, R. e Venter, J.C. (2013). A metagenomic framework for the study of airborne microbial communities. *PLOS one*, 8 (12), e81862.

Biclusterings

Adelaide Freitas, *adelaide@ua.pt*

Departamento de Matemática & CIDMA
Universidade de Aveiro

1. Motivação

O aumento do número de genomas sequenciados e a grande quantidade de dados complexos emergentes de tecnologias aplicadas ao ADN (ex: *microarrays*, *NGS-next generating sequences*) têm vindo a produzir novos desafios em vários domínios científicos nomeadamente em Estatística. Um desses desafios corresponde à identificação de padrões ou grupos homogéneos sobre dados genómicos. Por exemplo, em estudos de características da estrutura primária do genoma, a detecção de padrões semelhantes no contexto de pares de codões em genomas totalmente sequenciados, pode ser importante para desvendar as regras gerais que influenciam a fidelidade de descodificação do ARN-mensageiro [11, 12, 14]. Também na análise de dados de expressão de genes resultantes de experiências de *microarrays* de ADN, a identificação de genes com perfis de expressão semelhantes sob o mesmo subconjunto de condições experimentais é fundamental para a identificação de propriedades reguladoras de processos celulares [8].
A tecnologia dos *microarrays* permite a medição do nível de expressão de um grande número de genes sobre diferentes amostras ou condições experimentais. Os níveis observados de expressão dos genes são usualmente representados por uma matriz de dados numéricos X da forma:

Amostra *1*	x_{11}	x_{12}	...	x_{1J}
⋮	⋮	⋮	⋱	⋮
Amostra *I*	x_{I1}	x_{I2}	...	x_{IJ}
	Gene *1*	Gene *2*	...	Gene *J*

onde a observação na linha i e coluna j (x_{ij}) representa o nível de expressão do gene j na i-ésima condição experimental. O potencial de técnicas estatísticas de agrupamento (*clustering*) para revelar padrões biologicamente significativos foi inicialmente considerado por Eisen *et al.* (1998) [4], que aplicou agrupamento hierárquico para identificar grupos funcionais de genes sobre uma matriz de dados X. Contudo, as técnicas de agrupamento hierárquico, e outras tradicionais, não permitem que se sobreponham os grupos (*clusters*) pelo que as suas aplicações não se adequam a dados de sistemas biológicos em que o mesmo gene pode estar envolvido em múltiplos processos e, portanto, pertencer simultaneamente a vários grupos. Para superar algumas destas limitações, novas abordagens de agrupamento de dados foram propostas na última década [1, 2, 8, 9], nomeadamente na construção de algoritmos de agrupamento simultâneo de linhas (amostras) e colunas (genes) da matriz X para a identificação de subconjuntos de genes que estão co-expressos (*ie*., com perfil semelhante) sobre subconjuntos de amostras/condições da matriz X. Este tipo de agrupamento é chamado de *biclustering* [2] e os grupos resultantes são chamados de *biclusters*. Note-se que, nos *biclusterings,* os genes são agrupados de acordo com os níveis de expressão para algumas condições e não necessariamente para todas, como ocorre nas técnicas clássicas de *clusterings.* Em geral, uma aplicação de um algoritmo de *biclustering* determina quanto muito um *bicluster*. Várias aplicações podem detectar diferentes *biclusters* podendo um dado gene/condição pertencer a vários *bicluster*.

2. Definições

Formalmente, dada uma matriz de dados numéricos $X = [x_{ij}]$, $i \in \{1,2,\dots,I\}$, $j \in \{1,2,\dots,J\}$, de I indivíduos (ou amostras) e J variáveis (ou genes), um *bicluster* da matriz X é uma submatriz $[x_{ij}]$, $i \in S_i \subseteq \{1,2,\dots,I\}$, $j \in S_j \subseteq \{1,2,\dots,I\}$, onde as observações x_{ij} satisfazem alguma propriedade de grupo. Dependendo dessa propriedade, definem-se vários tipos de *biclusters* [13]. Por exemplo, se para todo o $i \in S_i, j \in S_j$ e k constante,

- $x_{ij} = k$, diz-se que $[x_{ij}]$ é um *bicluster de valores constantes* de X;
- $x_{ij} = k + a_i$ (modelo aditivo) ou $x_{ij} = k \times a_i$ (modelo multiplicativo), diz-se que $[x_{ij}]$ é um *bicluster de valores constantes em linha* de X sendo a_i o ajustamento para a linha i;
- $x_{ij} = k + b_j$ (modelo aditivo) ou $x_{ij} = k + b_j$ (modelo multiplicativo), diz-se que $[x_{ij}]$ é um *bicluster de valores constantes em coluna* de X sendo b_j o ajustamento para a coluna j;
- $x_{ij} = k + a_i + b_j$ (modelo aditivo) ou $x_{ij} = k \times a_i \times b_j$ (modelo multiplicativo), diz-se que $[x_{ij}]$ é um *bicluster de valores coerentes* de X;
- $x_{ij} = k \times a_i + b_j$ ou qualquer outra condição que traduza uma ordenação linear através das linhas (colunas, resp.), diz-se que $[x_{ij}]$ é um *bicluster de valores coerentes em linha* (*coluna,* resp.) de X.

Dependendo da estrutura do conjunto dos *biclusters* existentes numa matriz de dados, diferentes grupos de *biclusters* podem ser definidos [13]. Por exemplo, poder-se-ão ter *biclusters* exclusivos em linhas e/ou em colunas sempre que não existam linhas e/ou colunas comuns entre os diferentes *biclusters*; *biclusters* justapostos sempre que os *biclusters* são exclusivos em linha e/ou em colunas com a última linha ou coluna de um *bicluster* justaposta à primeira linha ou coluna de outro *bicluster*; *biclusters* sobrepostos sempre que os *biclusters* contiverem linhas ou colunas da matriz de dados que lhes sejam comuns.

Para ilustração, veja-se a Figura 1 onde para uma matriz (9×8), X, se identificam o *bicluster* B_1 de valores constantes em coluna, definido pela submatriz de linhas 1, 4, 9 e colunas 1, 4, 5, 6, 7, 8 de X, e o *bicluster* B_2 de valores constantes dado pela submatriz de linhas 1, 4, 6, 7, 9 e colunas 2, 5, 6 de X.

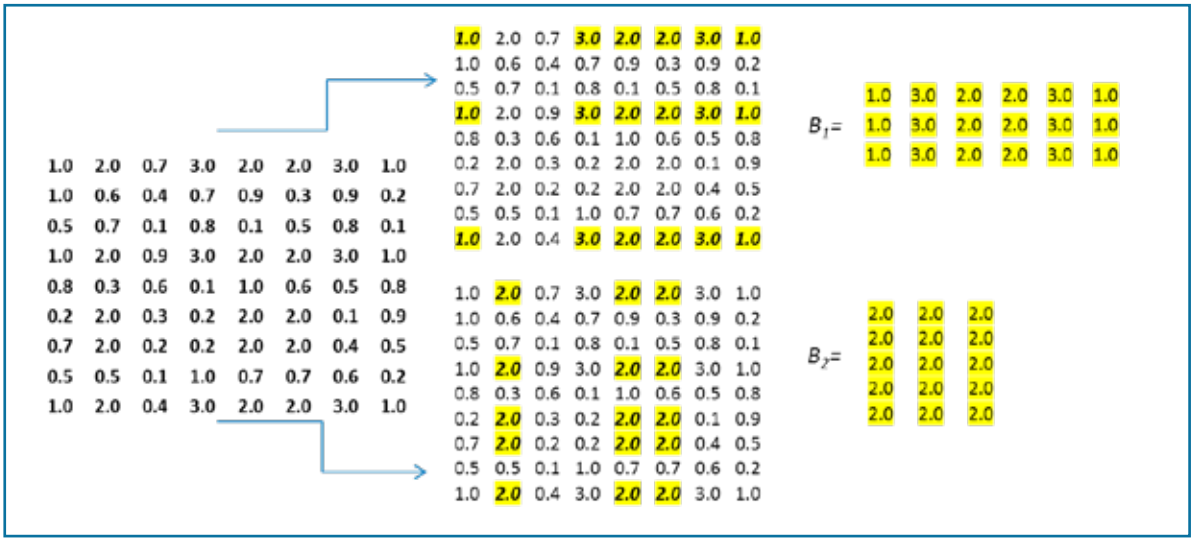


Figura 1. Ilustração de dois *bicluters* sobrepostos, B_1 com valores constantes em coluna e B_2 com valores constantes, identificados sobre uma matriz de dados (9×8).

Um *biclustering* é um algoritmo destinado a resolver um problema de optimização e corresponde à identificação do grupo de *biclusters* óptimos B_{opt} associados a X para os quais

$$f(B_{opt}) = \max_{B \in BC(X)} f(B)$$

onde f é uma função objectivo que mede a qualidade dos *biclusters* e $BC(X)$ é o conjunto de todos os possíveis grupos de *biclusters* associados a X. *Biclustering* é um problema combinatório num espaço de procura de tamanho $O(2^{I+J})$ sendo em geral NP-difícil [2, 6, 13]. Nenhum procedimento é ideal para encontrar o conjunto de *biclusters* óptimos. Um algoritmo de *biclustering* é definido por um

determinado critério de agrupamento simultâneo das linhas e das colunas da matriz e todo o algoritmo tem vantagens e desvantagens. Na avaliação do desempenho de um algoritmo sobre dados de expressão genética, tem particular interesse investigar a significância biológica dos *biclusters* detectados analisando, por exemplo, o nível de enriquecimento significativo dos genes envolvidos em cada *bicluster* comparativamente a redes biológicas conhecidas (por exemplo, as anotações na *Gene Ontology*, GO [7]) [18]. Em aplicações mais genéricas, a avaliação do desempenho passa por comparar a qualidade dos resultados usando funções que quantifiquem o grau de coerência dentro de cada *bicluster* e, na comparação entre diferentes técnicas de *biclusterings*, por recorrer a funções que meçam características comuns entre os grupos de *biclusters* identificados ou que quantifiquem a capacidade de um algoritmo revelar *biclusters* que tenham sido detectados por outro algoritmo [6, 10]. Em Freitas *et al.* (2013) [6] são detalhadas diversas funções de avaliação de algoritmos de *biclustering*.

3. Algoritmos de *biclustering*

Existem vários algoritmos de *biclustering*. Aqui são destacados o Bimax, aplicável a matrizes de dados binários, e o ISA que mostrou bom desempenho quer sobre dados simulados quer sobre conjuntos de dados de expressão genética [5, 15].

3.1.Bimax

O Bimax é um algoritmo de *biclustering* originalmente proposto por Prelic *et al*. (2006) [15] sobre matrizes de dados binários onde cada elemento (0 ou 1) define se não existe alteração (0) nos dados de expressão genética relativamente ao controlo ou se existe alteração (1). É baseado na estratégia *dividir & conquistar* e, contrariamente aos usuais métodos de *biclustering*, detecta todos os *biclusters* óptimos de uma matriz de dados binários. Um *bicluster* obtido usando o Bimax corresponde a uma submatriz com todos os elementos iguais.

3.2.ISA

O ISA (acrónimo de *Iterative Signature Algorithm*) é um algoritmo de *biclustering* originalmente proposto por Ihmels *et al*. (2002, 2004) [8, 9] para identificar, a partir de uma matriz de níveis de expressão genética, módulos de transcrição em experiências de *microarrays*. Um módulo de transcrição consiste de um conjunto de genes co-expressos associados a um conjunto de condições reguladoras de processos biológicos. Na sua versão original, o ISA é baseado no comportamento de médias ponderadas e é aplicado sobre matrizes com os genes posicionados em linha (*ie.*, usa a transposta de *X*). Em Freitas *et al.* (2011) [5] é proposto um algoritmo similar à estrutura do ISA tomando o comportamento da mediana em vez da média. Basicamente, o ISA foi desenvolvido para encontrar, numa matriz, submatrizes cujas linhas e colunas apresentem, simultaneamente, um comportamento, em média (ou em termos medianos), diferente do padrão *normal esperado*. O critério que define esse comportamento fora do padrão esperado para a média (ou mediana) das linhas e das colunas corresponde ao critério de agregação que determina os conjuntos de linhas e de colunas do *bicluster*. O padrão pode ser definido por forma a identificar as observações com valores mais elevados, as observações com valores mais baixos ou as observações com valores mais extremais (altos e baixos). Uma aplicação do ISA sobre uma matriz de dados origina quanto muito um único *bicluster*. Aplicando várias vezes, podem ser identificados múltiplos *biclusters*, sobrepostos ou não. Uma iteração do ISA é processada em dois passos (conhecido por algoritmo de assinatura). Na Figura 2 ilustra-se o procedimento geral do algoritmo de assinatura baseado nas médias para identificar *biclusters* com observações mais extremas. Inicialmente, o ISA parte de um subconjunto de linhas $L^{(0)}$(aleatoriamente seleccionadas ou não) e aplica iterativamente o algoritmo de assinatura até que algum critério de paragem seja satisfeito. Na primeira iteração, partindo de $L^{(0)}$, obtém-se um conjunto de colunas $C^{(0)}$ e um conjunto de linhas $L^{(1)}$. Na segunda iteração, obtém-se $C^{(1)}$, $L^{(2)}$, e assim sucessivamente. Um dos critérios de paragem habituais é duas iterações consecutivas produzirem o mesmo conjunto de linhas, $L^{(n+1)} = L^{(n)}$. O resultado final, a existir, corresponderá a um *bicluster* identificado pelos conjuntos de linhas $L^{(n)}$ e de colunas $C^{(n)}$.

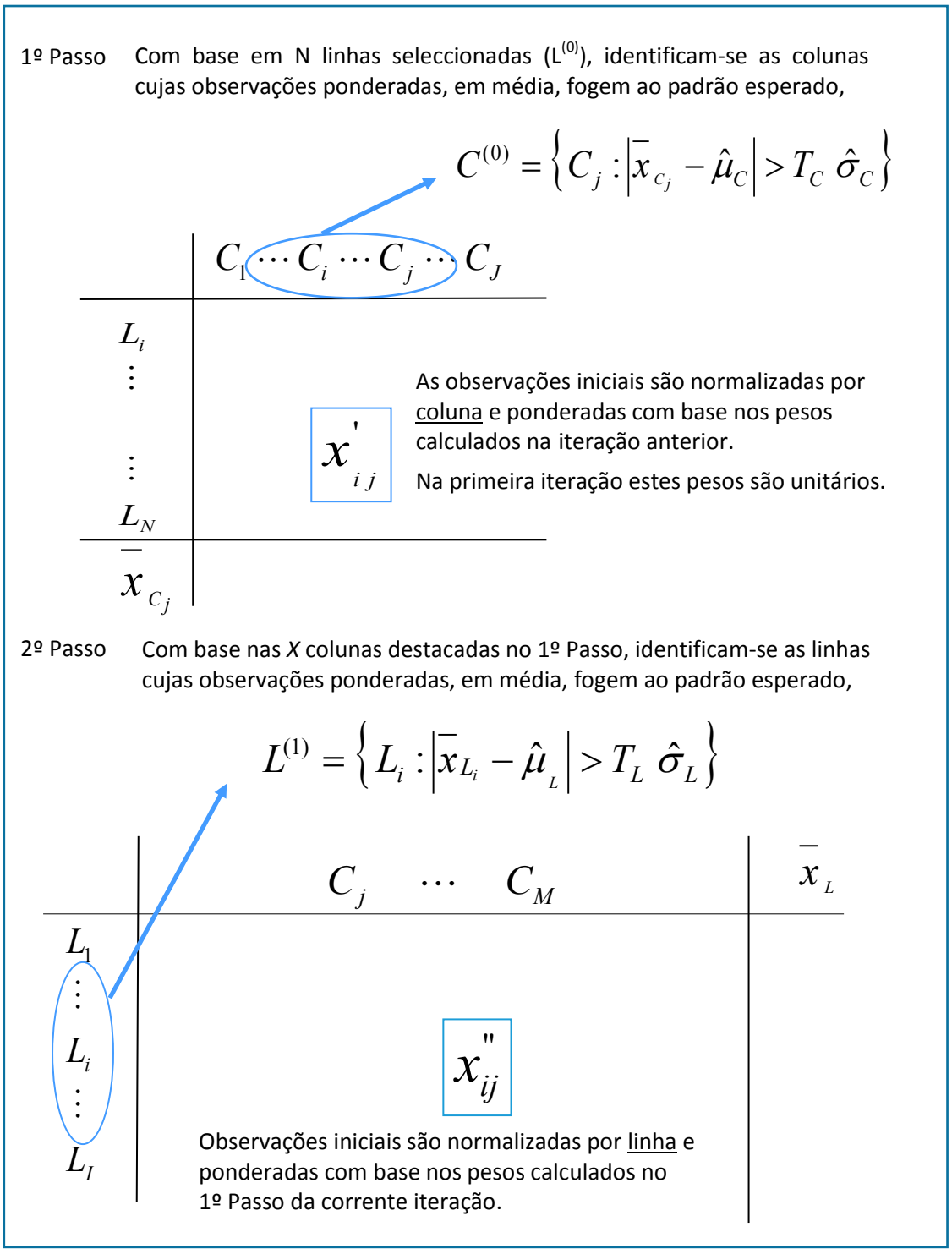


Figura 2. Algoritmo de Assinatura definido por uma iteração com dois passos. No primeiro (segundo) passo considera-se a matriz de dados normalizados por colunas (linhas). No primeiro passo, partindo de um conjunto de linhas escolhidas e de pontuações das linhas estabelecidas na iteração anterior, colunas cujas médias ponderadas não pertencem a um intervalo predefinido em termos de um parâmetro limiar T_C, são seleccionadas. No segundo passo, e para as colunas escolhidas no primeiro passo, o algoritmo selecciona todas as linhas cujas médias, ponderadas por pontuações das colunas estabelecidas no passo anterior, não pertencem a um intervalo predefinido em termos de um parâmetro limiar T_L. As pontuações em coluna (linha) são as médias ponderadas calculadas por coluna (linha) no passo (iteração) imediatamente anterior do algoritmo. Os pesos para o cálculo dessas médias ponderadas correspondem às pontuações em linha (coluna) obtidas na iteração (passo) imediatamente anterior do algoritmo.

4. *Biclusterings* usando o R

Vários algoritmos de *biclustering* encontram-se implementados no *software* estatístico R [16], nomeadamente, nas livrarias "`biclust`" [17] e "`isa2`" [3]. A primeira permite usar o Bimax e a segunda o ISA baseado na média.

Para ilustrar uma aplicação do Bimax, considere-se a matriz de dados binários resultante da matriz (9×8) da Figura 1 quando se transformam as observações x_{ij} em 0, quando $x_{ij} < 1$, e em 1, quando $x_{ij} \geq 1$. A seguinte sequência de comandos e de resultados do R permite identificar, quanto muito, 10 *biclusters*, com número de linhas e de colunas não inferiores a 3, obtidos por aplicação do Bimax sobre aquela matriz binária (`Xbinaria`) .

```
>install.packages("biclust")
>library("biclust")
>bics = biclust(x=Xbinaria, method=BCBimax(), minr=3, minc=3,
number=10)
>biclusternumber(bics)
$Bicluster1
$Bicluster1$Rows
[1] 1 4 6 7 9
$Bicluster1$Cols
[1] 2 5 6
$Bicluster2
$Bicluster2$Rows
[1] 1 4 9
$Bicluster2$Cols
[1] 1 2 4 5 6 7 8
```

São então identificados dois *biclusters*, o primeiro dado pela submatriz (5×3) de linhas 1, 4, 6, 7, 9 e colunas 2, 5, 6 de *X* e o segundo *bicluster* dado pela submatriz (3×7) de linhas 1, 4, 9 e colunas 1, 2, 4, 5, 6, 7, 8 de *X*.

Para ilustrar uma aplicação do ISA, considere-se a matriz (9×8) da Figura 1. A seguinte sequência de comandos do R permite obter os *biclusters* que apresentam perfis em linha e perfis em coluna elevados em média (`c("up", "up")`) e que são detectados em 100 aplicações do ISA sobre a matriz (`X`). Os parâmetros limiares considerados, na definição das fronteiras das regiões que determinam se os valores (normalizados) de *X* são mais elevados do que o esperado, correspondem a todas as possíveis combinações dos valores da sequência 0.2, 0.4, 0.6, 0.8, 1 para T_L (`thr.row`) e T_C (`thr.col`).

```
>install.packages("isa2")
>library("isa2")
>modules=isa(X,thr.row=seq(0.2,1,by=0.2),thr.col=seq(0.2,1,by=0.2),
+ no.seeds=100, direction=c("up", "up"))
```

Para visualizar os *biclusters* pode recorrer-se novamente ao pacote `biclust` procedendo previamente à conversão do objecto resultante do `isa2` num objecto do `biclust`. Concretamente, a seguinte sequência de comandos e de resultados do R permite identificar os *biclusters.*

```
> bc = isa.biclust(modules)
> biclusternumber(bc)
$Bicluster1
$Bicluster1$Rows
[1] 1 4 9
$Bicluster1$Cols
[1] 4 7
$Bicluster2
```

```
$Bicluster2$Rows
[1] 6 7
$Bicluster2$Cols
[1] 2 5 6
$Bicluster3
$Bicluster3$Rows
[1] 1 3 4 9
$Bicluster3$Cols
[1] 4 7
$Bicluster4
$Bicluster4$Rows
[1] 1 3 4 8 9
$Bicluster4$Cols
[1] 4 7
```

São então identificados quatro *biclusters*, o primeiro dado pela submatriz (3×2) de linhas 1, 4, 9 e colunas 4, 7 de *X*; o segundo dado pela submatriz (2×3) de linhas 6, 7 e colunas 2, 5, 6 de *X*; o terceiro dado pela submatriz (4×2) de linhas 1, 3, 4, 9 e colunas 4, 7 de *X*; o quarto dado pela submatriz (5×2) de linhas 1, 3, 4, 8, 9 e colunas 4, 7 de *X*. Para visualizar os *biclusters*, procede-se do seguinte modo:

```
> (GroupsBic=bicluster(X, bc))
$Bicluster1
     [,1] [,2]
[1,]    3    3
[2,]    3    3
[3,]    3    3
$Bicluster2
     [,1] [,2] [,3]
[1,]    2    2    2
[2,]    2    2    2
$Bicluster3
     [,1] [,2]
[1,]  3.0  3.0
[2,]  0.8  0.8
[3,]  3.0  3.0
[4,]  3.0  3.0
$Bicluster4
     [,1] [,2]
[1,]  3.0  3.0
[2,]  0.8  0.8
[3,]  3.0  3.0
[4,]  1.0  0.6
[5,]  3.0  3.0
```

Agradecimentos
Investigação parcialmente financiada por Fundos Nacionais através do CIDMA –Centro de Investigação & Desenvolvimento em Matemática e Aplicações– da Universidade de Aveiro e da FCT –Fundação para a Ciência e Tecnologia–, dentro do projecto UID/MAT/04106/2013.

NOTA: Este texto não foi escrito ao abrigo do novo Acordo Ortográfico.

Referências

[1] Barkow, S., Bleuler, S., Prelic, A., Zimmermann, P. e Zitzler, E. (2006) BicAT: a biclustering analysis toolbox. *Bioinformatics* 22 (2006), 1282-1283.

[2] Cheng, Y. e Church, G. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8, 93-103.

[3] Csardi, G., Kutalik, Z. e Bergmann, S (2010) Modular analysis of gene expression data with R. *Bioinformatics*. 26, 1376-7.

[4] Eisen, M. B., Spellman, P. T., Brown, P. e Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, 95, 14863-14868.

[5] Freitas, A., Afreixo, V., Pinheiro, M., Oliveira, J. L., Moura, G. e Santos, M. (2011) Improving the performance of the iterative signature algorithm for the identification of relevant patterns. *Statistical Analysis and Data Mining*, vol. 4, no. 1, pp. 71-83.

[6] Freitas, A., Ayadi, W., Elloumi, M., Oliveira, J. L. e Hao, J. K. (2013) A Survey on Biclustering of Gene Expression Data. In *Biological Knowledge Discovery Handbook: Preprocessing, Mining and Postprocessing of Biological Data*, E. Mourad and A. Zomaya Editors. Wiley Book Series on Bioinformatics: John Wiley & Sons, New Jersey, USA.

[7] Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25-29.

[8] Ihmels, J., Friedlander, G., Bergmann, S., Sarig,, O. e Ziv, Y. (2002) Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, 31, 370-377.

[9] Ihmels, J., Bergmann, S., e Barkai,N. (2004). Defining transcription modules using large-scale gene expression data. *Bioinformatics*, 13:1993-2003.

[10] Lui, X. e Wang, L. (2007) Computing the maximum similarity bi-clusters of gene expression data. *Bioinformatics*, 23, 50-56.

[11] Moura,G., Pinheiro,M., Silva,R., Miranda,I., Afreixo,V., Dias,G. Freitas,A., Oliveira,J.L., Santos, M. (2005) Comparative context analysis of codon pairs on an ORFeome scale. *Genome Biology*, 6, R28, 14 páginas.

[12] Moura G, Pinheiro M, Arrais J, Gomes AC, Carreto L, Freitas A, Oliveira JL, Santos M. (2007) Large Scale Comparative Codon-Pair Context Analysis Unveils General Rules that Fine-Tune Evolution of mRNA Primary Structure. *PLoS ONE* 2(9): e847, pp 10 páginas.

[13] Madeira, S. e Oliveira, A. (2004) Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 1, 24-45.

[14] Pinheiro, M., Afreixo, V., Moura, G., Freitas, A., Santos, M.A.S., Oliveira, J.L. (2006) Statistical, Computational and Visualization methodologies to unveil gene primary structure features. *Methods Inf Med,* 2, pp. 163-168.

[15] Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Bhlmann, P., Gruissem, W., Hennig, L.,Thiele, L. e Zitzler, E. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22, 1122-1129.

[16] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[17] Kaiser, S., Santamaria, R., Khamiakova, T., Sill, M., Theron, R., Quintales, L., Leisch, F. e De Troyer, E. (2015). biclust: BiCluster Algorithms. R package version 1.2.0. http://CRAN.R-project.org/package=biclust

[18] Tanay, A., Sharan, R. e Shamir, R. (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18 (Suppl. 1), S136-S144.

Meta-Análise de Dados de Transcritómica

José Caldas, *jose@insightomics.com*
Insightomics, Lda

Susana Vinga, *susanavinga@tecnico.ulisboa.pt*
IDMEC, Instituto Superior Técnico, Universidade de Lisboa

Introdução

A Meta-análise é o conjunto de métodos utilizados para comparar os resultados de vários estudos, com o objetivo de alcançar conclusões mais robustas (Rosenthal, 2001). No presente artigo, é feita uma revisão de abordagens de meta-análise para estudos de transcritómica, também conhecidos por estudos de expressão génica (*gene expression*), em biologia molecular.

O transcritoma é o conjunto de moléculas de ácido ribonucleico (ARN, *RNA - Ribonucleic Acid*) presentes numa população de células num dado instante. De uma forma simplificada, os genes, que são regiões específicas da molécula de ácido desoxirribonucleico (ADN, *DNA - Deoxyribonucleic Acid*) presente numa célula, são as unidades fundamentais da hereditariedade, levando à produção de moléculas de ARN, num processo designado por transcrição. Por sua vez, essas moléculas de ARN são utilizadas na produção de proteínas, num processo designado por tradução (Fig. 1). As proteínas têm uma variedade de funções, desde atividade enzimática à manutenção da estrutura celular. A regulação da produção de proteínas através do transcritoma é um processo dinâmico e complexo que envolve milhares de genes[1] e que é crítico para a sobrevivência da célula.

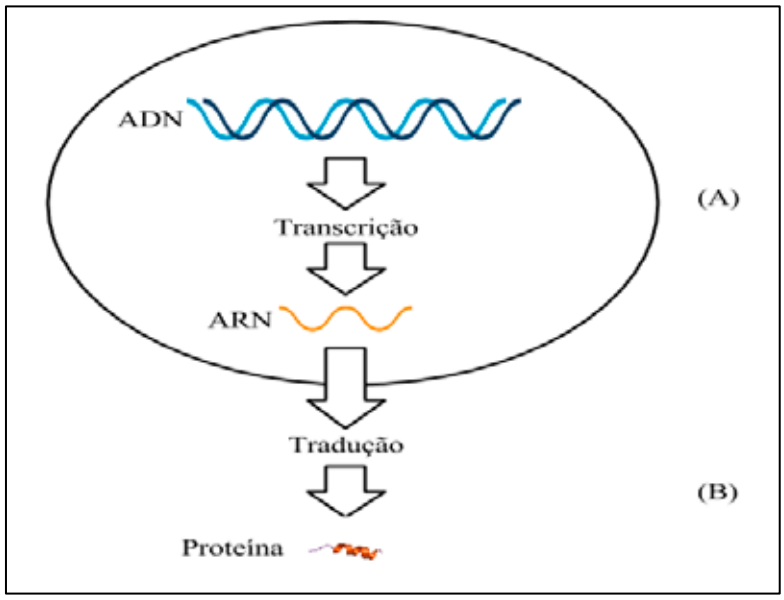


Fig. 1 - Ilustração simplificada dos processos de transcrição (A) e tradução (B) numa célula eucariótica. A transcrição ocorre no núcleo celular; a tradução ocorre no citoplasma.

[1] A título de exemplo, estima-se que, no ser humano, existam cerca de vinte mil genes (Ezkurdia, 2014).

As tecnologias laboratoriais principais para medir o transcritoma são os *microarrays* de ADN (Tarca, 2006), existentes desde o final dos anos noventa e, mais recentemente, técnicas de sequenciação de alto débito (RNA-sequencing ou *RNA-seq;* Wang, 2009). Estas tecnologias, ao permitirem caracterizar globalmente o transcritoma de uma população de células, tornaram-se ferramentas cruciais em áreas como a oncologia (Curtis, 2012) e a farmacologia (Lamb, 2006), mas introduziram também vários desafios a nível do processamento e análise de dados. O custo e complexidade de efetuar experiências de *microarrays* e *RNA-Seq* leva a que, tipicamente, um estudo inclua menos de dez amostras, embora cada amostra possa corresponder a milhares de medições de expressão génica. Este paradigma, conhecido por "*small n, large p*", diverge do paradigma clássico onde o número de amostras é superior à dimensionalidade de cada amostra, tendo levado à adaptação de várias metodologias de análise de dados (Johnstone e Titterington, 2009) e à criação de novas abordagens para o cálculo de significância estatística (Storey e Tibshirani, 2003).

Um dos problemas mais relevantes em análise de dados de transcritómica é, justamente, o da meta-análise. Devido à obrigatoriedade de armazenar, em bases de dados públicas, os dados de transcritómica gerados no contexto de um artigo científico, existem atualmente dezenas de milhares de experiências livremente disponíveis para análise (Parkinson, 2009). A possibilidade de reutilizar esses dados traz consigo não só o potencial de obter resultados robustos a partir de estudos relacionados, mas, também, o de encontrar ligações inesperadas que poderão permitir, por exemplo, a identificação de novas aplicações para fármacos já existentes. Nas seguintes secções efetuar-se-á uma revisão de abordagens recentes para a meta-análise de estudos de transcritómica.

Meta-Análise em transcritómica

Em termos formais, uma amostra de uma experiência de transcritómica pode ser representada por um vetor contendo as medições dos valores de expressão de cada gene. Assim sendo, os dados de um estudo com *n* genes e *m* amostras podem ser representados por uma matriz com *n* linhas e *m* colunas.
Muitos estudos de transcritómica baseiam-se em comparar amostras através de testes estatísticos, com o objetivo de detectar quais são os genes cujos padrões de expressão variam significativamente entre diferentes contextos biológicos (Rapaport, 2013). O resultado de um estudo deste tipo é uma lista com um valor *p* para cada gene. Neste contexto os testes mais habituais são o teste t e o teste de Mann-Whitney (Tarca, 2006).
Para efetuar a meta-análise dos valores *p* de um gene ao longo de vários estudos, é possível aplicar abordagens clássicas como o método de Fisher, que consiste em combinar os valores *p* de um gene segundo a fórmula:

$$x_g = \sum_{s=1}^{S} \log(p_{gs}), \quad (1)$$

em que p_{gs} é o valor *p* do gene *g* no estudo *s* e *S* é o número total de estudos. Sob a hipótese nula de que cada p_{gs} tem uma distribuição uniforme em [0, 1], x_g segue uma distribuição Chi-quadrado com *2S* graus de liberdade.
Na prática, um método relacionado com o método de Fisher, designado por *rank product*, foi originalmente proposto para testar se um gene está diferencialmente expresso entre diferentes condições, atingindo uma performance superior ao método de Fisher e outros em várias tarefas de meta-análise (Breitling, 2004; Hong e Breitling, 2008; Caldas e Vinga, 2014). Este método baseia-se em ordenar os genes de cada estudo de acordo com um critério como, por exemplo, o valor *p*, sendo a posição resultante do gene *g* no estudo *s* designada por r_{gs}.
De seguida, calcula-se, para cada gene, o produto das suas posições ao longo de todos os *S* estudos em análise:

$$r_g = \prod_{s=1}^{S} r_{gs}. \quad (2)$$

O produto r_g corresponde a uma pontuação global atribuída ao gene *g*. Para calcular a significância de r_g , considera-se o logaritmo de (2),

$$\log(r_g) = \sum_{s=1}^{S} \log(r_{gs}). \quad (3)$$

O valor *p* associado a (3) pode ser calculado usando uma abordagem semelhante à aplicada no método de Fisher, sob a hipótese nula H_0 de que, para um dado gene *g* e cada estudo *s*, r_{gs} tem uma distribuição discreta uniforme. Em detalhe, $u_{gs} = r_{gs} / (n + 1)$ tem uma distribuição aproximadamente uniforme em [0, 1], sendo *n* o número de genes no estudo. Aplicando o logaritmo a u_{gs}:

$$u_{gs} \sim \text{Unif}(0, 1), \quad (4)$$

$$-\log(u_{gs}) \sim \text{Exp}(1), \quad (5)$$

$$-\sum_{s=1}^{S} \log(u_{gs}) = -\sum_{s=1}^{S} \log(r_{gs}) + S\log(n+1) \sim \text{Gamma}(S, 1). \quad (6)$$

As equações acima podem ser utilizadas para calcular o valor *p* aproximado de r_g:

$$\begin{aligned} P(r_g \le x) &= P(\log(r_g) \le \log(x)) \\ &= P(-\log(r_g) \ge -\log(x)) \\ &= P(-\log(r_g) + S\log(n+1) \ge -\log(x) + S\log(n+1)) \\ &= P(\text{Gamma}(S, 1) \ge -\log(x) + S\log(n+1)). \end{aligned} \quad (7)$$

A distribuição Gama é usada para calcular o valor *p* aproximado de r_g. Na prática, esta abordagem é não só extremamente rápida mas também muito precisa (Koziol, 2010).

Uma abordagem alternativa consiste em usar diretamente a lista de genes considerados significativos em cada estudo, em vez dos seus valores *p*. Neste tipo de abordagem, testes estatísticos como o teste exato de Fisher permitem aferir se as listas de genes obtidas por dois estudos têm uma sobreposição significativa. Apesar da sua simplicidade, esta abordagem não é comum em meta-análise, devido ao facto da sobreposição entre listas de genes significativos ser tipicamente baixa, mesmo entre estudos semelhantes (Fan, 2006).
Para aumentar a reprodutibilidade e interpretabilidade de um estudo, é habitual mapear uma lista de genes diferencialmente expressos para conjuntos de genes previamente identificados, utilizando o teste de Fisher para quantificar a sobreposição entre a lista de genes significativos e cada um dos conjuntos de genes pré-definidos. Esses conjuntos de genes correspondem, por exemplo, a vias metabólicas ou genes envolvidos num dado processo biológico. Segal *et al*. utilizaram uma estratégia semelhante para efetuar uma meta-análise de estudos de cancro, obtendo os genes diferencialmente expressos em cada amostra e mapeando esses genes para conjuntos de genes pré-definidos (Segal, 2004). Esses resultados foram subsequentemente utilizados para derivar "módulos" de genes cuja ativação ou repressão está relacionada com o tipo de tumor.

Outra família de testes estatísticos para expressão diferencial evita a utilização de um limiar para separar os genes entre significativos ou não, reduzindo o número de parâmetros com que se tem de lidar (Subramanian, 2005). Estes métodos consistem em testar diretamente se um conjunto de genes está, no seu todo, diferencialmente expresso. O método mais popular nesta família, conhecido por *Gene Set Enrichment Analysis* (*GSEA*), permite detetar se um conjunto de genes atua em conjunto, apesar da expressão de cada um desses genes não ser individualmente significativa (Subramanian, 2005). Sucintamente, os genes num estudo são ordenados de acordo com um dado critério, por exemplo o seu valor *p* ou o rácio de expressão entre duas condições. Para um conjunto de genes *S*, calcula-se uma pontuação em cada posição da lista completa ordenada de genes,

$$x(i) = \sum_{j \le i : g_j \in S} \frac{1}{|S|} - \sum_{j \le i : g_j \notin S} \frac{1}{n - |S|}, \quad (8)$$

onde *i* designa uma posição na lista e *n* é o número total de genes no estudo. A pontuação final do conjunto de genes *S* é dada pelo máximo de $|x_i|$, sendo equivalente a uma estatística de Kolmogorov-

Smirnov. A significância dessa estatística pode ser obtida através de testes de permutação. Este método é ilustrado na Fig. 2.

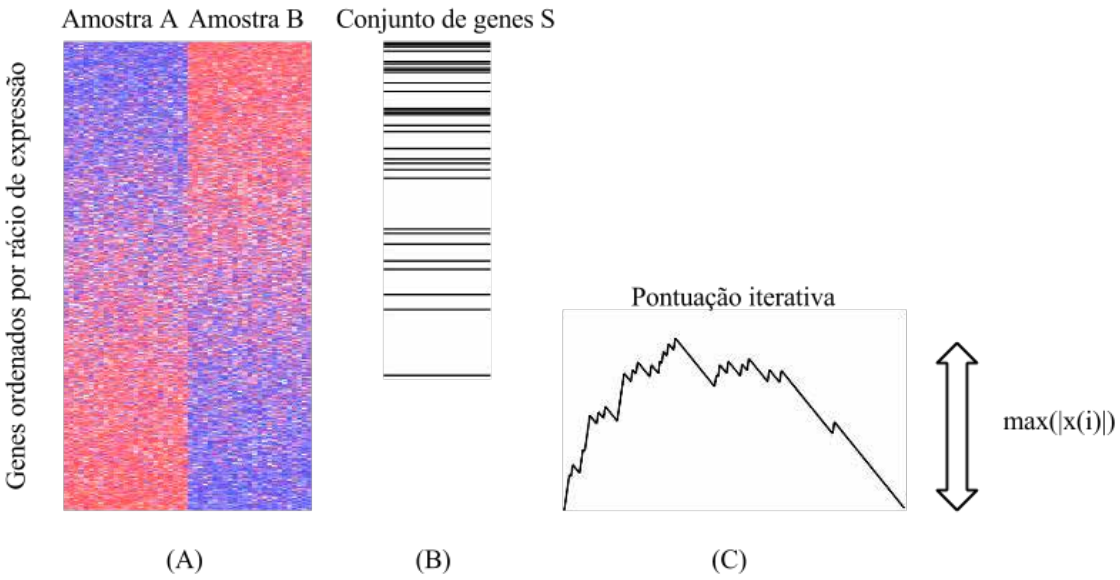


Fig. 2 - Método GSEA aplicado a dados de transcritómica envolvendo duas amostras A e B, adaptado de Subramanian *et al* (Subramanian, 2005). Em (A), todos os genes no estudo são ordenados de acordo com o rácio de expressão entre as amostras A e B. Em (B), os genes pertencentes ao conjunto *S* correspondem às linhas horizontais. Intuitivamente, grande parte dos genes em *S* aparece no topo da lista ordenada de genes. Em (C), a pontuação $x(i)$ é calculada iterativamente em cada posição da lista de genes. Esta pontuação aumenta quando um gene pertencente a *S* é encontrado, e diminui caso contrário. A estatística final corresponde ao máximo de $|x(i)|$.

A utilização deste método levou a uma maior reprodutibilidade entre estudos independentes de cancro do pulmão e está na base de várias abordagens de meta-análise (Subramanian, 2005). Estas abordagens consistem essencialmente em calcular, para cada estudo, os genes diferencialmente expressos (esta lista de genes é tipicamente designada de "assinatura" do estudo), e averiguar até que ponto a assinatura de um estudo aparece entre os genes mais ativos ou reprimidos de outros estudos, utilizando estatísticas semelhantes às usadas em *GSEA* (Kupershmidt, 2010). Este tipo de abordagem foi utilizado para identificar fármacos com potencial terapêutico para cancro do pulmão e doença inflamatória intestinal (Sirota, 2011).

As abordagens baseadas em testes estatísticos podem ser complementadas com outros métodos de estatística e aprendizagem automática, de forma a identificar padrões interessantes nos dados. Lukk *et al.* analisaram em conjunto cerca de cinco mil amostras de *microarrays* provenientes de vários estudos, mostrando, através de uma análise de componentes principais, que as amostras se organizam de acordo com o tipo de tecido e doença (Lukk, 2013). Caldas *et al.* combinaram o teste estatístico *GSEA* com um modelo de mistura, de forma a efectuar *clustering* de estudos e, ao mesmo tempo, obter os conjuntos de genes mais representativos de cada *cluster* (Caldas, 2009). Finalmente, Huang *et al.* utilizaram uma abordagem Bayesiana de classificação para obter um sistema de diagnóstico a partir de dados públicos de transcritómica (Huang, 2010).

Conclusões

No presente artigo fez-se uma revisão sucinta de métodos para meta-análise de dados de transcritómica. Um problema transversal a vários domínios é a combinação de estudos com pequenas amostras, que são tipicamente mais heterogéneos e levam a conclusões menos robustas (Haidich, 2010). Apesar de muitos estudos de transcritómica incluírem menos de uma dezena de amostras, o surgimento de grandes projetos de investigação, onde milhares de amostras são geradas (McLendon, 2008; Lamb, 2006), permitirá cada vez mais basear uma meta-análise em dados mais fiáveis.

Os vários estudos aqui apresentados demonstram que, apesar das dificuldades inerentes às tecnologias de transcritómica, é útil efetuar a meta-análise dos dados resultantes. Em geral, a meta-análise de estudos de transcritómica não é apenas um conjunto de metodologias para comparar resultados de vários estudos, mas, também, uma ferramenta de investigação de baixo custo em biologia molecular, que permite formular novas hipóteses em problemas relevantes, como identificar mecanismos moleculares de uma dada doença ou encontrar novas aplicações para fármacos existentes.

Agradecimentos

Agradecemos à FCT através do IDMEC, LAETA, projetos UID/EMS/50022/2013, Investigador FCT (IF/00653/2012) e CancerSys (EXPL/EMS-SIS/1954/2013).

Referências

Breitling, (2004). R. Breitling *et al*. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett*. 573:83-92, 2004.

Caldas, (2009). J. Caldas *et al*. Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics* 25:i145-153, 2009.

Caldas e Vinga, (2014) J. Caldas e S. Vinga. Global Meta-Analysis of Transcriptomics Studies. *PLoS ONE* 9:e89318, 2014.

Curtis, (2012) C. Curtis *et al*. The genomic and transcriptomic architecture of 2,000 breast tumors reveals novel subgroups. *Nature* 486:346-352, 2012.

Ezkurdia, (2014) I. Ezkurdia *et al*. Multiple evidence strands suggest that there may be as few as 19000 human protein-coding genes. *Hum. Mol. Genet.* 23:5866-5878, 2014.

Fan, (2006) C. Fan *et al*. Concordance among Gene-Expression–Based Predictors for Breast Cancer. *N. Engl. J. Med.* 355:560-569, 2006.

Haidich, (2010) A.B. Haidich. Meta-analysis in medical research. *Hippokratia* 14:29-37, 2010.

Hong e Breitling, (2008) F. Hong e R. Breitling. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics* 24:374-382, 2008.

Huang, (2010) H. Huang *et al*. Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *P. Natl. Acad. Sci. U.S.A.* 107:6823-6828, 2010.

Johnstone e Titterington, (2009) I. Johnstone e M. Titterington. Statistical challenges of high-dimensional data. *Philos. T. R. Soc. A.* 367:4237-4253, 2009.

Koziol, (2010) J. A. Koziol. Comments on the Rank Product Method for Analyzing Replicated Experiments. *FEBS Lett.* 584:941-944, 2010.

Kupershmidt, (2010) I. Kupershmidt *et al*. Ontology-based meta-analysis of global collections of high-throughput public data. *PLoS One* 5:e13066, 2010.

Lamb, (2006) J. Lamb *et al.* The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* 313:1929-1935, 2006.

Lukk, (2011) M. Lukk *et al*. A global map of human gene expression. *Nat. Biotechnol.* 28: 322–324, 2012.

McLendon, (2008) R. McLendon *et al*. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455:1061-1068, 2008.

Parkinson, (2009) H. Parkinson *et al.* ArrayExpress update — from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.* 37:D868-D872, 2009.

Rapaport, (2013) F. Rapaport *et al*. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* 14:R95, 2013.

Rosenthal, (2001) R. Rosenthal e M.R. DiMatteo. META-ANALYSIS: Recent Developments in Quantitative Methods for Literature Reviews. *Annu. Rev. Psychol.* 52:59-82, 2001.

Rung e Brazma, (2013) J. Rung e A. Brazma. Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.* 14:89-99, 2013.

Segal, (2004) E. Segal *et al*. A module map showing conditional activity of expression modules in cancer. *Nat. Genet.* 36:1090-1098, 2004).

Sirota, (2011) M. Sirota *et al*. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* 3:96ra77, 201.

Storey e Tibshirani, (2003) J. D. Storey e R. Tibshirani. Statistical significance for genomewide studies. *P. Natl. Acad. Sci. U.S.A.* 100:9440:9445, 2003.

Subramanian, (2005) A. Subramanian *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *P. Natl. Acad. Sci. U.S.A.* 102:15545-15550, 2005.

Tarca, (2006) A. L. Tarca *et al*. Analysis of microarray experiments of gene expression profiling. *Am. J. Obstet. Gynecol.* 195:373-388, 2006.

Wang, (2009) Z. Wang *et al*. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet*. 10:57-63, 2009.

Tudo sobre Malária, Genética, e Estatística, ou talvez não!

Nuno Sepúlveda, *nuno.sepulveda@lshtm.ac.uk*

London School of Hygiene and Tropical Medicine
Centro de Estatística e Aplicações da Universidade de Lisboa

Genética e Malária

Malária (ou paludismo) é uma doença predominantemente de climas tropicais causada pelos parasitas do género *Plasmodium – Plasmodium falciparum, Plasmodium vivax, Plasmodium ovale*, e por aí fora - que entram no organismo através de picadas dos mosquitos *Anopheles* infetados. O ciclo de infeção no homem inicia-se com a migração do parasita para o fígado onde ocorre uma multiplicação exacerbada do mesmo sem que os pacientes apresentem quaisquer sintomas. Esses parasitas são posteriormente redireccionados para a corrente sanguínea onde invadem os glóbulos vermelhos provocando a sua morte celular. É nesta fase que os doentes infetados mostram diversos sintomas clínicos, tais como, dores de cabeça, febre, calafrios, ou vómitos. Em casos extremos, os pacientes podem apresentar casos agudos de anemia por uma eliminação massiva de glóbulos vermelhos por parte do parasita. Os parasitas podem também atravessar a barreira hematoencefálica, provocando malária cerebral que pode conduzir à morte de um paciente. Neste cenário de largo espetro de sintomas, interessa perceber a razão pela qual existem indivíduos que parecem ser totalmente resistentes à infeção enquanto que outros têm uma predisposição natural para sofrer sintomas agudos da doença. A investigação em genética da malária tenta dar resposta a esta e a outras questões importantes para a saúde humana.

A importância de fatores genéticos na susceptibilidade à doença começa com a distribuição geográfica das diferentes espécies de *Plasmodium* pelo mundo. O *Plasmodium falciparum* predomina na África subsaariana enquanto que o *Plasmodium vivax* é o agente infeccioso mais importante na América do Sul e no Sudeste Asiático, estando praticamente ausente no continente africano. A razão para tal ausência é de natureza genética pois é sabido que, ao longo da evolução humana, as populações africanas adquiriram uma mutação particular no gene *Duffy* que impossibilita a invasão dos glóbulos vermelhos por parte dos parasitas *Plasmodium vivax*. Como contraponto, essa mesma mutação está praticamente ausente das populações sul americanas e asiáticas. Outro exemplo clássico da literatura é o gene associado à anemia falciforme. Do ponto de vista clínico, duas cópias deste gene provoca normalmente a morte do feto durante o período gestativo mas, em caso de sobrevivência do recém-nascido, conduz a uma condição aguda caraterizada por glóbulos vermelhos em forma de foice. Contudo, a presença de uma cópia apenas desse gene num individuo tende a conferir um efeito protetor contra sintomas agudos associados às infeções por *Plasmodium falciparum*. Esse efeito benéfico para o hospedeiro parece ter levado a uma seleção natural desse gene em África onde tal espécie é mais frequente. Conjuntamente ao gene da anemia falciforme, outros genes relacionados com a fisiologia dos glóbulos vermelhos assim como vários genes do sistema imunitário foram identificados como potenciais fatores protetores contra malária.

Há dez anos atrás o papel da genética em malária foi meticulosamente discutido na literatura especializada (Kwiatkovski, 2005). Foi identificado um problema de reprodutibilidade dos resultados genéticos derivado de vários fatores: desenhos amostrais inadequados, tamanhos amostrais reduzidos para atingir significância estatística, uso de diferentes marcadores genéticos, ausência de controlo de possíveis subpopulações geneticamente distintas presentes nos dados, e por aí fora. Na última década, muitos destes problemas foram colmatados com o avanço tecnológico em termos da sequenciação e catalogação das variantes genéticas presentes num determinado genoma. Esse avanço trouxe a oportunidade de os investigadores analisarem um elevado número de marcadores genéticos em tempo

real e a baixo custo num grande volume de amostras. Em particular, as tecnologias de sequenciação de nova geração permitiram explorar todo o genoma de forma a identificar vários tipos de variação genética simultaneamente, tais como, os populares SNPs, Indels, CNVs ou Inv (veja-se a Tabela 1 para as suas respetivas definições). Neste contexto, a genética da malária ganhou novo fôlego com o consórcio MalariaGEN que chamou a si uma vasta comunidade de investigadores espalhados um pouco por todo o mundo (http://www.malariagen.net). Atualmente este consórcio tem vindo a gerar 'terabytes' de informação genética à escala mundial, disponibilizando os respetivos dados na internet. É neste contexto que se faz uma breve revisão sobre este assunto, dando especial ênfase à variação genética do genoma humano e ao seu relacionamento com a malária, à variação genética dos parasitas da malária e ao seu relacionamento com a resistência a diferentes medicamentos anti-maláricos, e aos chamados estudos de associação genética.

Variante	Definição	Sigla (inglesa)
Polimorfismos num único nucleótido	Sequências de ADN que diferem entre si numa única posição	SNPs
Inserção-remoção	Inserções ou remoções de pequenas pedaços de ADN relativo a um genoma de referência	Indels
Alteração do número de cópias	Aumento ou diminuição do número de cópias de uma determinada sequência relativamente ao observado num genoma de referência	CNVs
Inversões	Inversão da sequência de ADN de uma determinada região em relação a um genoma de referência	Inv

Tabela 1: Definição das variantes genéticas mais comuns.

Variação genética do genoma humano e o seu relacionamento com a malária

Pode-se dizer que a sequenciação completa do genoma humano em 2001 foi uma dos grandes proezas científicas no início deste século. O genoma humano é composto por 23 pares de cromossomas - 22 pares de cromossomas autosomais e 1 par de cromossomas sexuais (X e Y) -, que contabilizam um total de 3,300 mega pares de bases (Mbp). Existem cerca de 21 mil genes já identificados mais outros tantos com o potencial de codificar alguma proteína. Desde da sua sequenciação foi criado um genoma de referência para estudos genéticos em populações humanas que, por sua vez, permitiu realizar mapas de variação genética humana. Na mesma linha de investigação de Mendel que usou sistemas de dois alelos para estudar a hereditariedade das ervilhas-de-cheiro, a variação genética mais simples de descobrir e de analisar é aquela relacionada com os SNPs. Em teoria, qualquer posição do genoma pode apresentar uma alteração no correspondente nucleótido. Contudo, em 2008, as estimativas para o número total de SNPs presentes no genoma humano não ia além dos 14 milhões; um catálogo de todos os SNPs até agora detetados pode ser encontrado no *website* do consórcio *hapmap* (http://hapmap.ncbi.nlm.nih.gov/). Note-se que estes SNPs podem resultar ou não numa alteração da sequência de aminoácidos da respetiva proteína. Tal como referido na introdução, o gene da anemia falciforme é um exemplo clássico de uma variante genética associado à malária. Esse 'gene' é de fato um SNP denominado por rs334, estando localizado no gene que codifica a hemoglobina B; este SNP implica a alteração do aminoácido Glutamato para Valina levando à produção de moléculas defeituosas dessa proteína. Existem ainda inversões e variações indels e em número de cópias, que eleva as estimativas da variação total para um patamar a rondar os 50 milhões de variantes genéticas. Um exemplo de uma inserção-remoção relacionada com malária é o do gene da talassémia alfa na hemoglobina A, que parece ter um efeito protetor contra infeções assintomáticas no nordeste da Tanzânia (Sepúlveda *et al*, 2014).

Importa sublinhar que as variantes genéticas presentes no genoma humano são o resultado de milhares de anos de evolução humana por resposta ao meio ambiente. Apesar de haver vários fatores ambientais já identificados (i.e., dieta ou fatores socioculturais), o parasita da malária teve um papel preponderante na evolução genética nas diferentes populações endémicas. Isso fica bem patente através de uma análise de componentes principais aplicada aos dados de vários SNPs em genes previamente associados com a malária (Figura 1A). Esta análise revela que a variação genética relacionada com esta doença parece ser suficiente para diferenciar as populações a nível continental. Para uma diferenciação de populações geograficamente próximas, é necessário entrar em linha de conta outras variações genéticas que, possivelmente, representem distintos processos evolutivos das respetivas populações. O problema da diferenciação das populações africanas é um tema bastante atual tendo sido criado um consórcio internacional para o efeito (Gurdasani *et al*, 2015). É sabido que o continente africano mostra uma diversidade genética mais elevada do que no resto do mundo mas as razões para tal observação ainda não são totalmente conhecidas. Tudo aponta para uma combinação de fatores étnicos com a presença de várias doenças infecciosas importantes, tais como é o caso da malária, febre de lassa, tripanosomíases, ou tracoma.

Variação genética do genoma dos parasitas *Plasmodium* e o seu relacionamento com a resistência a medicamentos anti-maláricos

Tal como foi referido na introdução, o género *Plasmodium* integra várias espécies. A mais estudada é o *Plasmodium falciparum* por ser o mais perigoso para a saúde humana. A sequenciação do seu genoma foi terminado em 2002 dando origem ao genoma de referência 3D7. Desde então várias atualizações têm sido avançadas e disponibilizadas na base de dados *plasmoDB* (http://plasmodb.org/). O comprimento total do genoma é de cerca de 23 Mbp distribuídos em 14 cromossomas, contabilizando perto de 5,300 genes. Tendo em conta que as máquinas de sequenciação da nova geração permitem sequenciar 20 a 120 Gbp numa única corrida, é atualmente possível obter informação detalhada de todo o genoma de um elevado número de amostras de parasitas. Apesar de todo esse potencial, o número de amostras analisadas em cada estudo é ainda reduzido devido a diversos problemas técnicos que incluem o armazenamento de grande volumes de dados em instituições académicas, ou a criação de procedimentos automáticos de análise que possam produzir resultados robustos para qualquer tipo de amostra.

Tal como no caso humano onde se usa um genoma de referência, o estudo da variação genética no *Plasmodium falciparum* é realizado em função do genoma de referência 3D7. Em geral, um genoma de um parasita proveniente de um paciente amostrado é alinhado com o de 3D7. O alinhamento de genomas é um problema estatístico uma vez que, em virtude da existência de variação genética no genoma a analisar, é preciso identificar o alinhamento mais provável de acordo com alguma métrica. Num trabalho recente, o alinhamento de 631 genomas de diferentes proveniências geográficas permitiu identificar um total de 600 mil SNPs no genoma do *Plasmodium falciparum*, sendo a maior parte deles de baixa frequência (Preston *et al*, 2014). Em consonância com o genoma humano, a análise de componente principais aplicada aos dados de SNPs permite também a identificação da origem geográfica dos parasitas *Plasmodium falciparum*, pelo menos a nível continental (Figura 1B). Este resultado demonstra que as populações destes parasitas são geneticamente distintas, tendo possivelmente sofrido diferentes mecanismos de diferenciação genética. Análises comparando as frequências alélicas entre populações através de estatísticas do tipo qui-quadrado permitiram identificar vários genes relacionados com a resistência a vários medicamentos anti-maláricos como fortes candidatos para essa diferenciação genética (Volkman *et al*, 2007). O exemplo mais conhecido da literatura é um conjunto de SNPs no gene *PFCRT* que confere ao parasita a capacidade de resistir a tratamentos baseados na cloroquina. Investigação está atualmente em curso para catalogar outro tipo de variações presentes no genoma deste parasita. Contudo, para que essa tarefa atinja uma elevada credibilidade científica, há a necessidade de desenvolver novos métodos de deteção de variação genética baseados em princípios fundamentais de análise estatística que rivalizem com os atuais métodos baseados em argumentos meramente heurísticos. Nessa linha de pensamento, um novo método estatístico para a deteção de CNVs foi desenvolvido e aplicado a dados genómicos do *Plasmodium falciparum* (Sepúlveda *et al*, 2013). De acordo com esse trabalho, foi possível identificar cerca de 1000 amplificações nas regiões codificantes do genoma, nomeadamente, uma no gene

PFMDR que parece ser responsável pela resistência a diversos medicamentos anti-maláricos atualmente presente no Sudeste Asiático.
Com respeito a outras espécies de *Plasmodium*, a catalogação da variação genética ainda está numa fase embrionária. Algumas estatísticas genómicas gerais podem ser encontradas na base de dados *plasmoDB* para diferentes espécies. Na mesma linha de investigação realizada para o genoma do *Plasmodium falciparum*, um trabalho recente demonstrou que também é possível identificar uma diferenciação continental nos parasitas *Plasmodium vivax* a partir de dados de SNPs (Baniecki *et a*l, 2015). Tal resultado carece ainda de uma explicação cabal mas porventura está relacionada com a pressão seletiva de diferentes medicamentos anti-maláricos usados em diferentes partes do mundo.

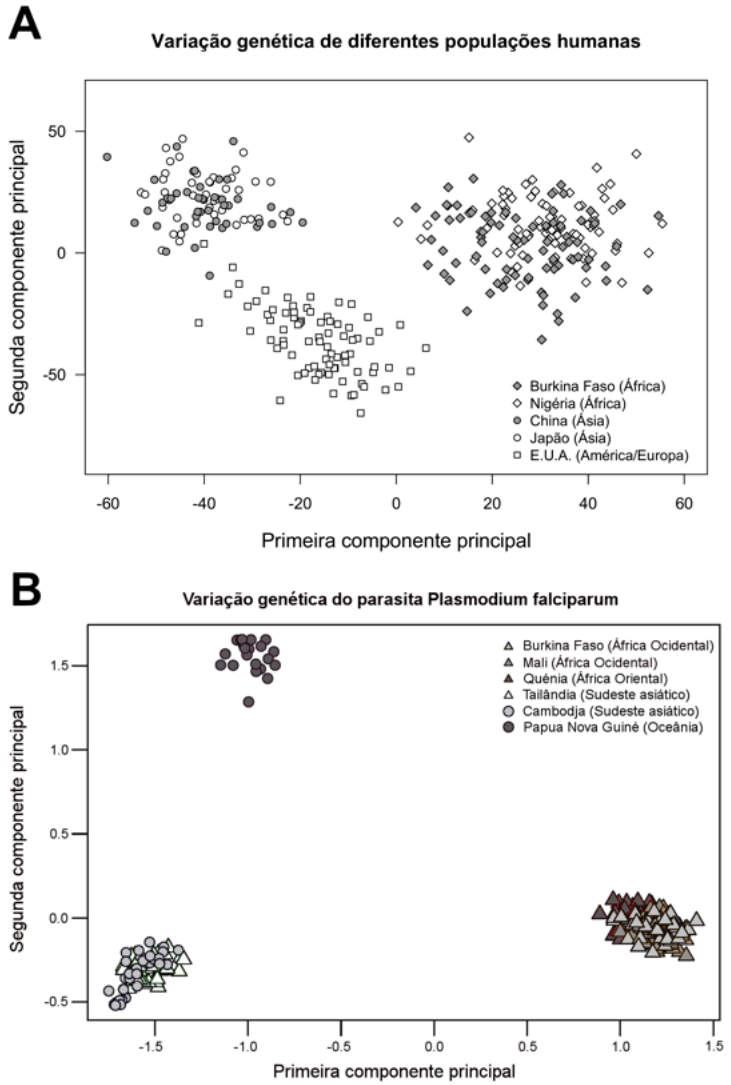


Figura 1: Variação genética de SNPs estudada por uma análise de componente principais. **A.** Dados de 163 SNPs em genes previamente associados com a resistência ou susceptibilidade à malária de várias populações humanas disponibilizadas no projeto HapMap (China – n=43, Japão – n=42, Europa/América – n=81, Nigéria – n=81) e de uma população do Burkina Faso (n=80, dados não publicados). **B.** Dados de cerca de 86 mil SNPs localizados ao longo de todo o genoma numa amostra de 227 parasitas de *Plasmodium falciparum* de diferentes proveniências geográficas (gráfico adaptado de Manske et al, 2012).

Estudos de associação genética

Um dos principais objetivos da Genética consiste em descobrir os mecanismos de hereditariedade que expliquem a variação de uma determinada caraterística biológica observável – vulgarmente conhecido por fenótipo – num determinado momento evolutivo de uma espécie. No cerne desses mecanismos está o conjunto de genes que contribuem para tal variação e a sua interação com o meio ambiente no sentido mais lato do termo. Deste ponto de vista, um geneticista necessita de identificar em primeiro lugar os genes potencialmente responsáveis pelas variações observadas no fenótipo, para depois desvendar os mecanismos biológicos subjacentes.

Os chamados estudos de associação genética visam descobrir os locais do genoma associados com um fenótipo. A ideia geral consiste em avaliar a associação estatística entre uma variação genética com a mesma observada ao nível do fenótipo. Para esse fim, existem vários desenhos experimentais, tais como, os estudos em famílias que permitem controlar possíveis efeitos de subestruturas genéticas na população mas de difícil recrutamento de todos os membros das famílias amostradas, estudos caso-controlo para fenótipos binários (e.g., resistência ou susceptibilidade a uma determinada doença), ou estudos do tipo transversal onde se amostra da população um conjunto de indivíduos potencialmente independentes entre si. Em qualquer destes estudos é normal assentar-se a análise em dados de SNPs pelo seu maior conhecimento genómico e pela sua maior facilidade em determinar o seu genótipo. Os SNPs a analisar podem ser apenas aqueles localizados em genes hipoteticamente responsáveis pelo fenótipo (estudos em genes candidatos) ou, então, aqueles disponibilizados em plataformas experimentais do tipo *microarray* que contemplam milhares ou até milhões de SNPs espalhados ao longo do genoma (estudos GWA – *genome wide association*). Em termos da averiguação estatística da associação genética propriamente dita, é usual analisar-se um SNP de cada vez e executar-se um teste de hipóteses que avalie a sua significância num determinado modelo estatístico. A título ilustrativo, se se pretende analisar um fenótipo binário, é comum usar-se modelos de regressão logística onde se aplica um teste da razão de verosimilhanças para comparar um modelo com o efeito do SNP em análise com um outro sem esse mesmo efeito. A significância do primeiro modelo em relação ao segundo dá ideia de quão forte é a associação entre um SNP e o fenótipo. Contudo, como um determinado estudo pode contemplar um elevado número de SNPs, tal como é o caso dos estudos GWA, existe um problema de testes múltiplos no mesmo conjunto de dados. Para colmatar esse problema, há que ajustar o nível significância de cada teste individual em função do nível de significância global que se pretende atingir na análise. No caso dos estudos GWA, esse ajustamento pode ir até valores entre 10^{-5} e 10^{-7}, dependendo do número de SNPs em análise, o que implica um tamanho de amostra extremamente elevado para a deteção de alguma associação genética estatisticamente significativa.

Na última década testemunhou-se uma proliferação de vários estudos de associação genética em malária. Muitos deles foram impulsionados pela criação do consórcio MalariaGEN. Neste consórcio existe um centro de recursos que disponibiliza aos seus membros não só um moderno serviço de sequenciação mas também uma equipa de estatísticos especializados em dados genéticos, que serve de força motriz à publicação dos respetivos resultados científicos. Até agora, a maior parte dos estudos de associação genética centrou-se na componente humana da doença. Os fenótipos mais estudados são aqueles relacionados com a apresentação ou não de sintomas clínicos agudos (Jallow *et al*, 2009; Manjurano *et al*, 2012; Toure *et al*, 2012; Timmann *et al*, 2012), ou relacionados com a susceptibilidade ou resistência à infeção (da Silva Santos *et al*, 2012; Maiga *et al*, 2014). A título ilustrativo, explora-se um estudo realizado no nordeste da Tanzânia (Sepúlveda *et al*, 2014), que não cai diretamente nos estudos padrões de associação genética. Este estudo contemplou uma amostragem transversal de mais 8000 indivíduos assintomáticos distribuídos em 24 aldeias localizadas em altitudes entre 165m e 1872m. Note-se que a altitude é aqui considerada uma possível medida do grau de exposição à malária, uma vez que as condições óptimas para a transmissão da doença decrescem com essa variável. A análise centrou-se em 175 SNPs localizados em diferentes genes previamente descritos como potenciais reguladores da maior ou menor resistência à doença. O objetivo da análise foi avaliar a variação genética desses SNPs em função de medidas de intensidade de malária refletindo efeitos de curto, médio e longo prazo de exposição à doença (prevalência atual de infeção, seroprevalência aos parasitas da malária e altitude, respetivamente). Aqui apresenta-se a associação genética relativa a efeitos de longo prazo (*e.g.*, altitude). Com esse fim, as distribuições genotípicas de cada SNP foram calculadas em cada aldeia e analisadas como dados composicionais usando a transformação log-aditiva e modelos multivariados de regressão linear. Figura 2 sumaria os respetivos resultados da associação entre cada um dos marcadores genéticos e altitude. Constata-se que apenas três associações são estatisticamente significativas usando uma correção de Bonferroni para o problema de testes múltiplos. A associação mais forte provém da conhecida inserção-remoção no gene da talassémia alfa. Uma associação intermédia é a do popular SNP rs334 localizado no gene da hemoglobina B e que dá origem à anemia falciforme. Uma associação na fronteira de significância estatística é um SNP no gene CD36. Estas três associações são, então, o ponto de partida para uma análise mais detalhada dos genes identificados, tendo em vista o refinamento da localização da

variante genética verdadeiramente associada com altitude. Isto pode ser feito pela sequenciação de um maior conjunto de SNPs nos genes detetados, ou então usar métodos de imputação de dados para inferir os genótipos de SNPs não contemplados na análise mas que se sabe serem altamente correlacionados com os SNPs identificados, tal como ilustrado em Jallow *et al* (2009) para os SNPs localizados no gene da hemoglobina B.

Até muito recentemente, os estudos de associação genética do ponto de vista do parasita foram realizados num conjunto muito reduzido de genes. Com o advento das modernas tecnologias de sequenciação, dois estudos GWA tentaram descobrir os genes do *Plasmodium falciparum* envolvidos com a atual resistência a diferentes medicamentos anti-maláricos (Borrmann *et al*, 2013; Takala-Harrison *et al*, 2013). Estes estudos redireccionaram a mesma metodologia estatística usada em estudos GWA humanos mas com algumas adaptações devido à especificidade dos respetivos dados. Na prática, a resistência aos medicamentos anti-maláricos é normalmente determinada por dois métodos: (1) ensaios de inibição de crescimento em culturas de parasitas onde se calcula o valor IC50 referente à dose necessária para atingir 50% da densidade de parasitas obtida a partir de uma dose de referência; (2) estimativa do tempo que um determinado medicamento leva a reduzir 50% da carga parasitária inicial num paciente seguido ao longo do tempo. Por um lado, as estimativas do valor IC50 não seguem tipicamente distribuições gaussianas e, por isso, há a necessidade de aplicar uma transformação conveniente aos dados para que isso aconteça. Outra solução é usar-se testes não paramétricos para estudar a associação genética com eventual perda de potência estatística (Borrman *et al*, 2013). Por outro lado, os tempos até à redução de 50% da carga parasitária também não apresentam formas distribucionais gaussianas. Neste caso, uma solução simples mas que acarreta possivelmente uma perda de informação estatística consiste em dividir os tempos em 'lento' e 'não-lento' usando um valor de corte com relevância clínica (Takala-Harrison *et al*, 2013). Assim, transforma-se um fenótipo quantitativo contínuo num binário onde se pode usar modelos de regressão logística. Ao contrário dos estudos genéticos em populações humanas onde o genótipo de um SNP é determinado inequivocamente, salvo eventuais erros de sequenciação, vários estudos genéticos nos parasitas *Plasmodium falciparum* demonstraram a possibilidade de se obter genótipos múltiplos para alguns SNPs devido ao fato do mesmo paciente poder apresentar infeções recorrentes. Essa eventualidade passa a ser uma realidade quando se estuda populações com alta incidência de malária. Uma forma de contornar este problema consiste em concentrar a análise nos SNPs que não apresentem quaisquer genótipos múltiplos. Contudo, essa solução pode reduzir drasticamente o número de SNPs em análise, diminuindo assim a probabilidade de identificar a verdadeira variante genética responsável pela variação fenotípica. Uma solução mais elegante consiste em tratar esses genótipos múltiplos como dados omissos e analisa-los como tal por algum método desenhado para esse tipo de dados. Sendo possível na prática, esta solução carece ainda de uma fundamentação teórica que assegure uma elevada qualidade estatística dos respetivos resultados.

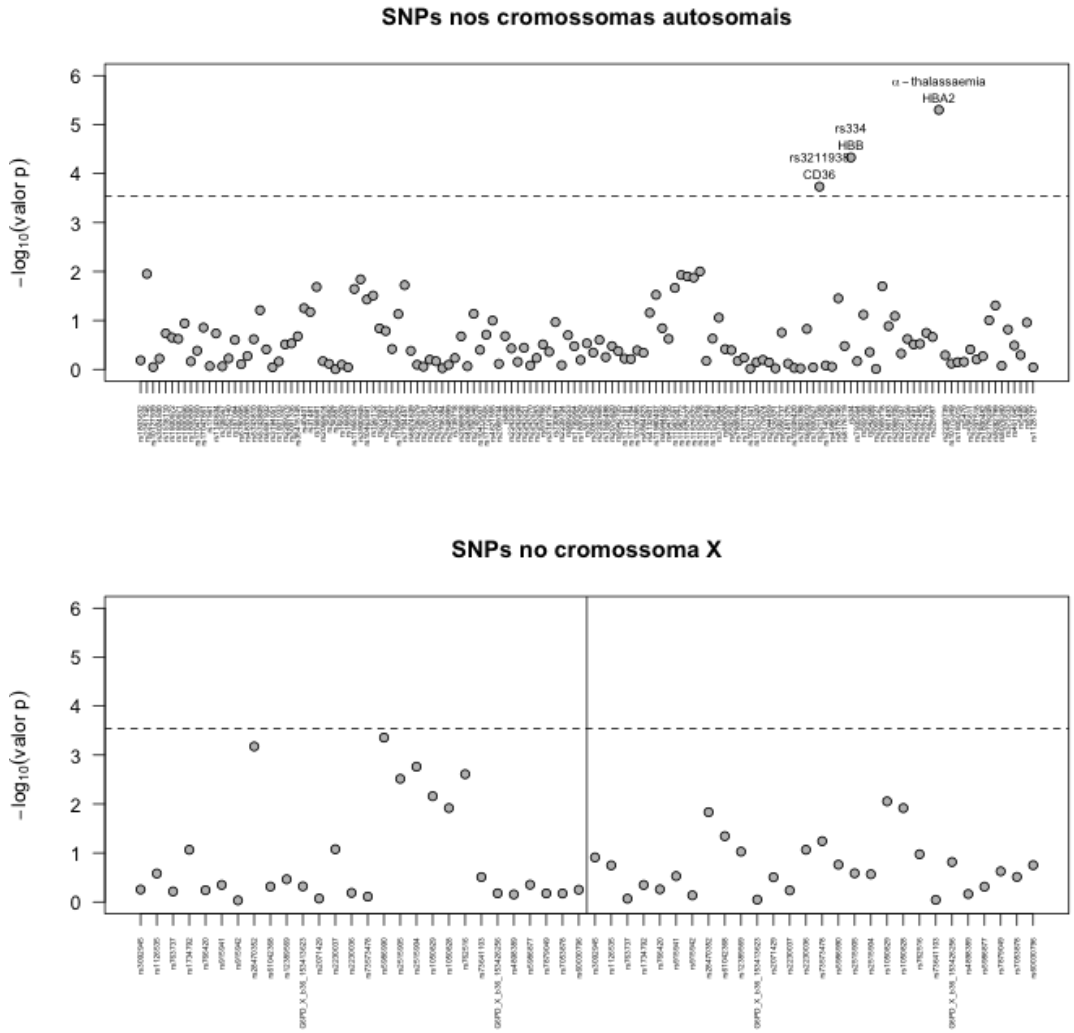


Figura 2: Força da associação entre SNPs localizados nos cromossomas autosomais e X e altitude em dados provenientes de um estudo realizado no nordeste da Tanzânia, onde o valor p é referente ao teste da razão de verosimilhanças testando o efeito de altitude em explicar a variação das distribuições genotípicas de cada aldeia e de cada marcador genético, e o valor de corte para a significância estatística de um determinado marcador (linha a tracejado) foi derivado da correção de Bonferroni para testes múltiplos de forma a garantir um nível de significância global de 5%.

Algumas considerações finais

Tendo a certeza que muito ficou por escrever, espera-se que esta breve revisão estimule um interesse da comunidade estatística de língua portuguesa sobre este assunto. Por um lado, os atuais dados genéticos usados em investigação da malária são gerados pelas modernas tecnologias de sequenciação, estando na vanguarda do progresso científico. Assim sendo, a união da estatística à malária mostra ser uma aliança estratégica para afirmar o valor da disciplina nesta nova era de grande volume de dados. Por outro lado, com o acumular de dados genéticos associados a malária e não só, a investigação em estatística aplicada tem luz verde para tentar várias soluções de forma a extrair a maior informação possível destes conjuntos de dados teoricamente ricos. Em última análise, o desenvolvimento de metodologia estatística para a análise genética de malária pode trazer benefícios para outros campos científicos com problemas similares.

Uma questão que tem sido sucessivamente posta de lado ao longo dos anos é a da interação genética. Ao contrário dos fenótipos mendelianos onde existe apenas um único gene em ação, os fenótipos normalmente estudados em malária estão sob a influência de múltiplos genes, tal como ficou patente na Figura 2 para o estudo do nordeste da Tanzânia. É então pertinente perguntar como estes diferentes genes interagem entre si na construção do fenótipo e qual o seu possível efeito na performance estatística dos estudos de associação genética que analisam um SNP de cada vez. Esta questão foi discutida nalguns trabalhos publicados no âmbito da SPE para o caso de dois genes (revisto em Sepúlveda, 2009), mas ainda não abordada no presente contexto de um maior conhecimento dos genomas humano e dos parasitas da malária. O futuro parece então oferecer uma oportunidade única

de conceber ferramentas estatísticas que permitam inferir a arquitetura genética da resistência à malária de forma a usá-la no desenvolvimento de vacinas eficazes contra esta doença (Kwiatkovski, 2005).

Agradecimentos
Agradeço ao Prof. Taane Clark, Dr. Kevin Tetteh, e Dr. Khalid Beshir da London School of Hygiene and Tropical Medicine (LSHTM) por alguns esclarecimentos sobre a literatura da malária.
Agradeço também ao Prof. Chris Drakeley pelo apoio demonstrado nos últimos anos que me permitiu apreciar o mundo da malária com um todo.

Referências

Baniecki, M. L., *et al*. (2015). Development of a Single Nucleotide Polymorphism Barcode to Genotype *Plasmodium vivax* Infetions. *PLoS Negl. Trop. Dis.*, vol. 9, e0003539.

Borrmann, S., *et al*. (2013). Genome-wide screen identifies new candidate genes associated with artemisinin susceptibility in *Plasmodium falciparum* in Kenya. *Sci. Rep.*, vol. 3, 3318.

da Silva Santos, S., *et al* (2012). Investigation of host candidate malaria-associated risk/protetive SNPs in a Brazilian Amazonian population. *PLoS One*, vol. 7, e36692.

Gurdasani, D., *et al* (2015). The African Genome Variation Projet shapes medical genetics in Africa. *Nature*, vol. 517, 327–332.

Jallow, M., *et al* (2009). Genome-wide and fine-resolution association analysis of malaria in West África. *Nat. Genet.*, vol. 41, 657-665.

Kwiatkovski, D. P. (2005). How Malaria Has affeted the Human Genome and What Human Genetics Can Teach Us about Malaria. *Am. J. Hum. Genet.*, vol. 77, 171-192.

Manske, M., *et al* (2012). Analysis of *Plasmodium falciparum* diversity in natural infetions by deep sequencing. *Nature*, vol. 487, 375-379.

Maiga, B. *et al* (2013). Human candidate polymorphisms in sympatric ethnic groups differing in malaria susceptibility in Mali. *PLoS One*, vol. 8, e75675.

Manjurano A, *et al* (2012). Candidate human genetic polymorphisms and severe malaria in a Tanzanian population. *PLoS One*, vol. 7, e47463.

Preston, M. D., *et al* (2014). PlasmoView: A Web-based Resource to Visualise Global *Plasmodium falciparum* Genomic Variation. *J. Infet. Dis.*, vol. 209, 1808-1815.

Sepúlveda N., *et al* (2014). On the performance of multiple imputation based on chained equations in tackling missing genotypes in a malaria association study in Tanzania. *Ann. Hum. Genet.,* vol. 78, 277-89.

Sepúlveda, N., *et al* (2013). A Poisson hierarchical modelling approach to deteting copy number variation in sequence coverage data. *BMC Genomics*, vol. 14, 128.

Sepúlveda, N. (2009). Abordagem por penetrâncias alélicas: uma nova abordagem estatística para análise de fenótipos binários complexos. Em *Estatistica: Arte de explicar o acaso* (Oliveira, I., Correia, E., Ferreira, F., Dias, S., Braummann, C., editores), p. 109-129, Edições SPE, Lisboa.

Takala-Harrison, S., *et al* (2013). Genetic loci associated with delayed clearance of Plasmodium falciparum following artemisinin treatment in Southeast Asia. Proc. Natl. Acad. Sci. U.S.A., vol. 110, 240-245.

Timmann, C., *et al* (2012). Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature*, vol. 489, 443-446.

Toure, O. *et al* (2012). Candidate polymorphisms and severe malaria in a Malian population. *PLoS One*, vol. 7, e43987.

Volkman, S. K., *et al* (2007). A genome-wide map of diversity in *Plasmodium falciparum*. *Nat. Genet.,* vol. 39, 113-119

Integração de informação biológica num modelo Bayesiano hierárquico para deteção de variantes genéticas

Miguel Pereira, *miguel.pereira14@imperial.ac.uk*

National Heart and Lung Institute, Imperial College London, Reino Unido

1. Introdução

Os estudos de associação genética têm como objetivo investigar a base genética de fenótipos mensuráveis através do estudo da relação entre variantes genéticas e doenças ou outras características mensuráveis (ex. altura, índice de massa corporal)[1](referência 1). As variantes genéticas habitualmente estudadas são os polimorfismos de nucleótido único, mais conhecidos por *single nucleotide polymorphisms* (SNPs), que são variações na sequência do DNA que afetam apenas um nucleótido e constituem a principal fonte de variação genética entre indivíduos. A identificação de variantes associadas a uma fenótipo pode ser efetuada de duas formas: através do estudo de SNPs localizados num pequeno conjunto de genes que se pensa estarem associados ao fenótipo (estudo de genes candidatos) ou através do estudo de SNPs em todo o genoma (estudo de associação *genome-wide* – *genome-wide association study* ou GWAS).

Um GWAS envolve o estudo de milhões de SNPs em milhares de indivíduos nos quais é observada a característica fenotípica ou doença e, habitualmente, o número de SNPs, p, excede largamente o número de indivíduos, n. A abordagem estatística clássica de análise de dados de GWAS baseia-se na estimação individual do efeito de cada SNP numa análise de regressão linear ou logística com a definição de um limiar de significância muito estrito (em geral, $valor - p < 5 \times 10^{-8}$) para controlar o erro tipo I e tomar em consideração os testes múltiplos, pois são realizados p testes. Com esta abordagem, poucos SNPs "sobrevivem" a este limiar e os SNPs detetados explicam apenas uma pequena proporção da heritabilidade predita para a doença, o que constitui o problema da 'heritabilidade em falta"[2]. Fundamentalmente, esta abordagem SNP-a-SNP tem pouco poder estatístico para detetar verdadeiras associações tanto no caso de variantes genéticas frequentes na população mas com pequenos efeitos, como no caso de variantes genéticas raras e com uma grande magnitude de efeito. Uma forma de resolver este problema de poder estatístico, é aumentar o tamanho da amostra, n, o que tem sido conseguido através da criação de grandes consórcios de colaboração que recrutam e sequenciam milhares de indivíduos. Esta medida contribuiu para a deteção de mais variantes associadas com doenças, no entanto, ainda estamos longe de resolver o problema da "heritabilidade em falta". Adicionalmente, o aumento do tamanho da amostra não é viável em todas as situações, como o caso de fenótipos difíceis de medir por implicarem medidas invasivas, uma logística complexa ou custos elevados.

A análise de SNPs em conjunto, isto é, análise de regressão que inclui todos os SNPs numa única equação, tem sido sugerida como uma forma de melhorar a deteção de SNPs em GWAS. A exploração da matriz de correlação entre SNPs, permite obter estimativas mais precisas dos efeitos de cada SNP[3]. Adicionalmente, a análise em conjunto permite estudar interações entre SNPs para além dos efeitos individuais de cada variante. Apesar desta abordagem ser claramente vantajosa, a sua implementação tem sido limitada. Tal como já referido, num GWAS o número de SNPs, p, excede largamente o

número de indivíduos, n, o que resulta num problema de $p >> n$ que impede a estimação dos efeitos de cada SNP utilizando métodos clássicos como o método dos mínimos quadrados. Têm sido desenvolvidos vários métodos que lidam com este problema, nomeadamente em casos de dados de grandes dimensões [4], e alguns dos quais têm sido aplicados à análise de dados de genómica[5,6]. No entanto, o número de SNPs investigados é na ordem das várias centenas de milhar a milhões o que impede a implementação destes métodos de forma computacionalmente eficiente.

O Desequilíbrio de Ligação e dados de grande dimensão

Um dos problemas da análise estatística clássica em GWAS é o pressuposto de independência entre os vários SNPs que permite a análise de cada SNP individualmente. No entanto, SNPs muito próximos entre si no mesmo cromossoma apresentam uma tendência para serem herdados em conjunto, um fenómeno conhecido como Desequilíbrio de Ligação (DL), também conhecido por desequilíbrio de *linkage*, e que se traduz em correlação estatística entre SNPs. O DL é medido através da comparação da frequência da combinação de SNPs em dois *loci* próximos com a frequência esperada assumindo que os SNPs são independentes.

Tendo em conta a existência desta estrutura de correlação e aplicando medidas de desequilíbrio de ligação, é possível dividir o genoma em blocos de DL que traduzem a dependência entre SNPs numa população[7]. A importância de considerar estes blocos na análise de GWAS advém de dois argumentos: primeiro, a correlação entre SNPs pode ser explorada na análise conjunta para melhorar a estimação dos efeitos de cada variante, e segundo, a exploração desta estrutura de correlação pode ser utilizada para reduzir a dimensionalidade nos GWAS. A eficiência computacional tem impedido a implementação de muitos modelos de análise conjunta de GWAS devido ao problema da dimensionalidade dos dados. Este problema é ainda maior em modelos Bayesianos, que são, em geral, mais exigentes computacionalmente. A consideração da estrutura de blocos de DL pode permitir a redução de dimensionalidade quer através da utilização dos blocos como unidade de análise ao invés das variantes genéticas, quer através da implementação de métodos de estimação que têm em conta o agrupamento de SNPs por bloco e fazem a estimação por grupo sendo, consequentemente, mais eficazes computacionalmente.

Alguns estudos exploraram a estrutura de blocos de DL para aumentar o número de variantes detetadas em GWAS e que ilustram o potencial de utilizar esta estrutura para melhorar a deteção de SNPs. Em particular, Tregouët *et al*. utilizaram blocos de DL como unidade de análise para identificar um novo *locus* associado a doença coronária[8]. Ehret *et al*. usaram também uma estrutura de grupos de SNPs e verificaram uma melhoria significativa, em relação à abordagem clássica, na proporção de variância explicada pelos SNPs detectados em associação com a altura, índice de massa corporal e razão anca-altura[9]. Mais recentemente, Pare *et al*. estudaram SNPs agregados em grandes regiões cromossómicas e conseguiram identificar novos SNPs com pequenos efeitos localizados junto de *loci* com associação conhecida[10].

Informação Biológica Externa e Modelos Bayesianos em Estudos de Associação Genética

Outra forma de melhorar a deteção de SNPs em GWAS é a integração de informação biológica externa sobre cada SNP na análise estatística. Atualmente, há uma panóplia de informação genómica disponível em várias bases de dados *online* que pode ser obtida automaticamente e integrada na análise. Para este fim, a abordagem Bayesiana representa uma escolha óbvia, dado que a inclusão de informação externa é uma parte importante da sua estrutura, permitindo a integração de forma muito flexível de múltiplos tipos de evidência externa. O racional por trás desta abordagem baseia-se na hipótese de que a detecção de SNPs pode ser melhorada através de atribuição de maior peso a SNPs que apresentam um papel biológico, penalizando SNPs que não aparentam ter qualquer função.

Vários modelos Bayesianos têm sido aplicados na análise de estudos de associação genética. Em geral, estes modelos incluem métodos genéricos que lidam com o problema de $p >> n$, onde os métodos de estimação clássicos não podem ser utilizados. Apesar de vários autores sugerirem a inclusão de

informação externa na análise, a literatura onde é efetuada a integração desta informação é ainda escassa, principalmente no contexto dos GWAS. A integração de informação biológica tem sido aplicada na análise de dados de expressão genética através da inclusão de informação de vias biológicas, redes de genes e fatores de transcrição, tendo-se verificado uma melhoria da detecção de expressão diferencial. Detalhes sobre a inclusão de informação externa neste contexto encontram-se largamente descritos em Vannucci *et al*. 2011[11].

O objetivo deste trabalho é demonstrar uma abordagem em desenvolvimento com o propósito de otimizar a detecção de SNPs em GWAS através da inclusão de informação biológica externa. Esta informação é incluída num modelo Bayesiano que efetua análise conjunta de SNPs agrupados em blocos de DL. Espera-se que esta abordagem aumente o poder estatístico de detecção de variantes genéticas comparativamente à atual abordagem de análise individual de SNPs (abordagem clássica), fazendo uso da grande quantidade de informação biológica *a priori* disponível.

2. Métodos

Esta secção consiste na descrição dos métodos utilizados para obter e incluir informação biológica externa (2.1) num modelo Bayesiano que estima os efeitos de SNPs em conjunto (2.2). O benefício da inclusão da informação biológica é avaliado num conjunto de dados do *European Respiratory Community Health Survey* (ECRHS) (2.3).

2.1 Obtenção de informação biológica externa

O primeiro aspeto considerado na obtenção de informação biológica externa foi a seleção do tipo de informação a incluir no modelo estatístico, isto é, a informação que influencia a magnitude de associação de um SNP com um fenótipo. Foi incluída informação potencialmente relevante relativa a dez 10 itens/perguntas (Figura 1) baseado no trabalho de Minelli *et al*.[12]. Neste trabalho, os autores investigaram a importância de catorze tipos de informação biológica prévia relativa ao SNP ou gene/região genómica onde o SNP se localiza e demostraram que a inclusão desta informação melhora a priorização de SNPs em sete fenótipos/doenças distintos. Quatro dos itens utilizados pelos autores não foram utilizados neste trabalho visto dependerem de literatura prévia, podendo enviesar os resultados, pois essa informação vai favorecer *loci* previamente estudados, penalizando *loci* nunca identificados por apresentarem efeitos mais difíceis de detetar, ou seja, aqueles que se pretende identificar com os novos métodos estatísticos.

A informação relativa aos 10 itens/perguntas (Figura 1) foi obtida através do *Data Integrator* ("*Dintor*"), uma ferramenta bioinformática desenvolvida na European Academy of Bolzano (EURAC, Itália) que obtém informação de forma automatizada de múltiplas bases de dados de genómica *online*. Foi ainda obtida informação de outras bases de dados: *Pfam Protein Family Database*, para as questões 3 e 4, *Mouse Genome Informatics* (MGI, *http://www.informatics.jax.org*), para a questão 8 e *Reactome* (*www.reactome.org*) para a questão 10.

Questões de informação biológica *a priori*

1. O SNP encontra-se numa região transcrita mas não traduzida?
2. O SNP encontra-se numa região traduzida mas não altera o aminoácido?
3. O SNP altera o aminoácido de uma proteína?
4. O SNP encontra-se numa zona funcional da proteína?
5. O SNP encontra-se numa região regulatória que não é transcrita?
6. O SNP encontra-se numa região regulatória que é transcrita?
7. O SNP encontra-se numa região genómica conservada em vertebrados?
8. O SNP encontra-se num gene (±5kb)* associado ao mesmo fenótipo/fenótipo semelhante em modelos funcionais (com animais ou *in vitro*)?
9. O SNP encontra-se num gene (±5kb)* altamente expresso num tecido relevante para o fenótipo?
10. O SNP encontra-se num gene (±5kb)* que apresenta interação genéticas ou proteicas relevante para o fenótipo?

Figura 1 – Lista de itens de informação biológica externa incluída no modelo estatístico.
**Foram incluídos todos os genes num intervalo de ±5kb do SNP.*

2.2 Modelo Bayesiano

Foi utilizado o método desenvolvido por Yi *et al*.[13], um modelo linear generalizado Bayesiano com distribuições *a priori* de encolhimento (*shrinkage*) que efetua análise conjunta de variáveis divididas em grupos. Este modelo tem várias vantagens que motivaram a sua utilização. Primeiro, efetua a análise conjunta de milhares de SNPs de forma computacionalmente eficaz através de uma modificação do algoritmo Iterativo dos Mínimos Quadrados Ponderados (IWLS) que adiciona os passos Expectativa-Maximização (EM). Segundo, permite incorporar uma estrutura de grupo na análise, neste caso a estrutura de blocos de DL. Por fim, através de modelação de alguns hiperparâmetros, permite incluir informação biológica. Este modelo encontra-se implementado no pacote do R *BhGLM* e pode ser acedido em http://www.ssg.uab.edu/bhglm.

O método baseia-se num modelo Bayesiano hierárquico com a seguinte formulação:

$$\alpha_j|\tau^2_{\alpha_j} \sim N(\mu_j, \tau^2_{\alpha_j})$$
$$\tau^2_{\alpha_j}|s^2_{\alpha_j} \sim Inv-\chi^2(1, s^2_{\alpha_j})$$
$$s^2_{\alpha_j}|b_{k[j]} \sim Gamma(a, b_{k[j]})$$
$$p(\log b_k) \propto 1$$

onde α_j corresponde ao coeficiente do *j*-ésimo SNP, μ_j o valor inicial para o efeito do SNP (aqui definiu-se sempre $\mu_j = 0$, assumindo-se que o SNP não está associado ao fenótipo), $\tau^2_{\alpha_j}$ corresponde à variância da distribuição *a priori* dos efeitos dos SNPs e $s^2_{\alpha_j}$ corresponde a um hiperparâmetro de escala. $b_{k[j]}$ corresponde a um hiperparâmetro de escala específico de grupo atribuindo diferentes graus de encolhimento a cada grupo para além do encolhimento específico de cada SNP definido por $s^2_{\alpha_j}$.

Este modelo permite ainda a estimação dos efeitos de cada grupo para além dos efeitos dos SNPs ("Modelo Completo"). No entanto, adoptou-se o modelo mais simples onde são apenas estimados efeitos dos SNPs dado que a informação *a priori* é ao nível do destes e não ao nível dos blocos de DL e porque o "Modelo Completo" apresenta tendencialmente mais falsos positivos[13].

2.3 Teste do modelo em dados da ECRHS

O modelo desenvolvido por Yi *et al*.[13] foi testado usando dados relativos ao Índice de Massa Corporal (IMC) em indivíduos incluídos no estudo ECRHS, um estudo europeu iniciado no anos Oitenta para

estudar fatores associados ao aumento mundial da prevalência de asma[14]. O IMC foi escolhido visto ser um fenótipo largamente estudado em vários GWAS, existindo vários SNPs identificados e replicados em pelo menos um estudo. Usando dados relativos a 1829 indivíduos e 2.588.592 SNPs, foi construída uma base de dados com um número inferior de SNPs com o objetivo de reduzir o tempo computacional. Para construção da base de dados, foram selecionados SNPs que se sabe estarem associados ao IMC ("verdadeiros" SNPs) para testar a performance do modelo na discriminação de verdadeiros SNPs e dos blocos de DL onde estes se localizam.

A seleção dos verdadeiros SNPs teve como base uma meta-análise recente de GWAS relativos ao IMC que identificou 97 SNPs significativos[15]. Utilizando os dados do estudo ECRHS, foi efetuada a análise estatística clássica SNP-a-SNP para cada um destes SNPs, bem como para todos os SNPs nos mesmo blocos de DL e foram selecionados 6 SNPs verdadeiros de acordo com os seguintes critérios: 3 SNPs altamente significativos, 2 SNPs com valor-p (relativo ao efeito do SNP) próximo de 0.05 e 1 SNP com um valor-p próximo de 0.1. Todas as regressões foram ajustadas para a idade, sexo, centro onde foram efetuadas as medições e três componentes principais relativos à origem ancestral dos indivíduos de forma a controlar para eventual estratificação populacional, um fator de confundimento frequente estudos de genética populacional. Adicionalmente, foram selecionados aleatoriamente 24 SNPs a partir dos 2,588,592 SNPs da base de dados completa que foram usados como controlos ("falsos" SNPs). Os trinta SNPs (6 "verdadeiros" e 24 "falsos") foram mapeados nos respetivos blocos de DL utilizando a ferramenta *Pos2LDBlock* implementada no *Dintor*[16]. Com esta abordagem obteve-se um conjunto de 2614 SNPs e 1829 indivíduos para testar a performance do modelo.

A inclusão de informação biológica externa teve como base a modelação dos hiperparâmetros $s^2_{\alpha_j}$ e $b_{k[j]}$ que controlam o encolhimento ao nível do SNP e do grupo (bloco de DL), respetivamente. Dado que o efeito (e direção do efeito) é desconhecido para todos os SNPs assumiu-se, por defeito, que $\mu_j = 0, j = 1,2,\dots,p$, onde p corresponde ao número total de SNPs. A inclusão de informação biológica prévia teve como base o seguinte racional: aos SNPs com mais informação biológica prévia aplica-se menor encolhimento à volta de μ_j de forma a aumentar a probabilidade de α_j ser diferente de zero. Pelo contrário, aos SNPs com pouca ou sem informação biológica externa aplica-se maior encolhimento de forma a obter, com maior probabilidade, efeitos nulos. Tendo como base este princípio, e sabendo que fazendo variar os parâmetros $s^2_{\alpha_j}$ e $b_{k[j]}$ entre 0.1 e 2.5 se obtém um intervalo de encolhimento de forte a fraco, respetivamente, foi efetuada uma correspondência linear entre o número de perguntas biológicas com resposta "sim" (Figura 1) e os parâmetros $s^2_{\alpha_j}$ e $b_{k[j]}$ (que se fizeram variar entre 0.1 e 2.5). Esta escolha deveu-se à escassez de literatura relativa à forma de inclusão de informação prévia num modelo Bayesiano deste tipo, desconhecendo-se a melhor forma de realizar este tipo de correspondência.

3. Resultados

Para estudar a eficácia do método e o efeito da inclusão de informação biológica externa, foram testados seis tipos de modelos, três sem inclusão de informação biológica e três com inclusão desta informação. Estes modelos foram comparados com os resultados obtidos na análise estatística clássica. Os modelos sem inclusão de informação biológica externa foram os seguintes: (1) análise conjunta dos SNPs sem consideração da estrutura de blocos de DL, (2) análise conjunta de SNPs tendo em conta a estrutura de blocos e com encolhimento constante aplicado localmente em cada bloco, (3) análise conjunta de SNPs tendo em conta a estrutura de blocos com encolhimento variável específico para cada grupo. A inclusão de informação biológica foi efetuada de três formas distintas: (4) modelação de $s^2_{\alpha_j}$ atribuindo-se um valor distinto para cada SNP com base nas respostas às perguntas, (5) modelação de $s^2_{\alpha_j}$ atribuindo-se o mesmo valor aos SNPs de cada grupo com base na média das respostas de cada

SNP dentro de um bloco e (6) modelação de $s^2_{\alpha_j}$ atribuindo-se o mesmo valor aos SNPs de cada grupo com base no máximo de respostas "sim" obtidas por um SNP em cada bloco.
Para todos os modelos obteve-se estimativas dos efeitos dos SNPs, intervalos de credibilidade e valores-p Bayesianos. Os SNPs foram ordenados de acordo com o seu valor-p Bayesiano e a performance dos vários modelos foi avaliada tendo em conta o ranking do SNP com menor valor-p de cada grupo, critério que foi utilizado para ordenar os blocos de DL. Os resultados do ranking dos 30 blocos de DL em cada um dos modelos utilizados encontram-se na Figura 2. Dentro de cada tipo de modelo, foram testados vários hiperparâmetros das distribuições *a priori*, tendo-se considerado os cenários de encolhimento forte, moderado e fraco fazendo variar os hiperparâmetros $s^2_{\alpha_j}$ e $b_{k[j]}$. Os modelos apresentados correspondem a cenários de encolhimento moderado.

	Modelos						
		Sem Inclusão de Informação Biológica			Com Inclusão de Informação Biológica		
		1	**2**	**3**	**4**	**5**	**6**
	Análise Clássica	Análise conjunta sem grupos	Encolhimento por grupo, constante	Encolhimento específico por grupo	Encolhimento específico por SNP	Encolhimento específico por grupo (média)	Encolhimento específico por grupo
1	22	3	3	17	19	16	16
2	29	13	13	5	24	25	25
3	19	22	24	25	3	8	3
4	17	17	22	1	22	23	24
5	28	16	25	13	25	18	30
6	13	25	17	23	17	5	17
7	20	24	7	3	13	29	9
8	12	9	30	9	7	3	6
9	9	7	19	21	28	24	23
10	30	30	28	29	30	17	8
11	6	5	16	11	16	6	18
12	21	28	5	4	5	11	5
13	3	19	9	7	29	9	22
14	8	23	23	19	20	22	13
15	24	20	29	20	6	19	20
16	10	29	18	15	2	30	29
17	4	18	20	2	23	12	19
18	7	12	6	22	18	26	28
19	25	6	12	28	8	13	4
20	18	11	1	27	4	21	11
21	2	1	11	26	14	20	7
22	16	4	2	24	12	28	12
23	14	26	26	30	26	15	2
24	5	2	4	6	11	2	26
25	1	8	8	16	9	4	21
26	11	14	14	18	15	7	15
27	23	15	15	14	1	1	10
28	15	21	21	8	21	10	1
29	27	10	10	10	10	27	14
30	26	27	27	12	27	14	27

Figura 2 - Ranking dos blocos de DL de acordo com o SNP de cada bloco com melhor ranking em cada modelo. Os blocos de DL que contêm SNP que se sabe estarem associados encontram-se sombreados a cinzento. Os números correspondem a identificadores para cada bloco de DL baseado no cromossoma e posição.

A abordagem clássica classificou 4 dos 6 blocos de DL verdadeiros no top 5 sendo que o pior bloco de DL verdadeiro ficou classificado na 19ª posição, o que é coerente com a escolha previamente efetuada dos "verdadeiros" blocos de acordo com os resultados da análise clássica. Os resultados do modelo 1, onde a estrutura de grupo não foi considerada e o grau de encolhimento é transversal a todos os SNPs, sugerem que existe alguma melhoria da análise conjunta. Os verdadeiros blocos de DL encontram-se no top 16, o que é sugestivo da utilidade da análise conjunto para aumentar o poder estatístico para detetar associações. O modelo 2 apresenta o melhor ranking dos modelos sem informação biológica prévia. Os 6 blocos de DL "verdadeiros" encontram-se nos 15 melhores SNPs, com 3 blocos no top 6. Os resultados do modelo 2 são ainda consistentes com os resultados do modelo 1 com os blocos de DL mais altos no ranking no modelo 1 mantendo-se altos no ranking no modelo 2, verificando-se o mesmo com os blocos pior classificados.
O modelo 3 apresentou a pior performance, com bloco de DL pior classificado na 22ª posição. Observa-se que os "verdadeiros" blocos de DL encontram-se espalhados pelo ranking e que existe menor consistência comparativamente com os outros modelos. Isto sugere que a estimação, a partir dos dados, de um parâmetro de encolhimento específico de grupo pode ser afetada por características específicas dos blocos de DL, tais como o número de SNPs em cada bloco e a correlação entre SNPs em cada bloco, que prejudicam a estimação.
Dos modelos com inclusão de informação biológica prévia é de destacar o modelo 4 onde se observa uma clara melhoria relativa aos modelos sem informação biológica. Quatro dos verdadeiros blocos encontram-se no top 6 e o bloco de DL "verdadeiro" pior classificado encontra-se na 13ª posição. Os modelos 5 e 6 são controversos com subida da classificação de alguns blocos "verdadeiros" em detrimento de outros blocos de DL "verdadeiros" que apresentaram pior classificação. É de salientar que os blocos 24 e 25, subiram significativamente de posição o que ilustra a potencial utilidade de usar informação biológica externa para detectar blocos que não teriam sido detectados na análise clássica.

4. Discussão

O presente trabalho pretende demonstrar alguns resultados que ilustram o efeito da análise conjunta de SNPs e, principalmente, a inclusão de informação externa de dados de genómica num modelo Bayesiano.
A aquisição de informação biológica constitui um passo fundamental neste trabalho que envolve utilização de ferramentas de bioinformática que obtêm dados de múltiplas bases de dados, a tradução desta informação de forma a ter um significado relativo à associação entre o SNP e a doença/fenótipo e a sua inclusão num modelo estatístico. Neste trabalho, foi obtida informação tendo como base um conjunto de perguntas de resposta binária e biologicamente relevantes. A informação das perguntas foi utilizada para modelar o encolhimento aplicado ao coeficiente de cada SNP variando assim a probabilidade de detetar cada variante. Foi considerada a possibilidade de incluir a informação biológica através da modelação do efeito *a priori* do SNP dado por μ_j, o que seria uma forma mais direta e intuitiva de incluir a informação externa. No entanto, a modelação de μ_j implica "escolher" a direção do efeito (positiva ou negativa) de cada SNP, o que não é conhecido *a priori* visto que os itens de informação externa são se debruçam sobre a direção do efeito. Outra possibilidade considerada foi a inclusão da informação biológica modelando a probabilidade de um SNP estar associado com o fenótipo à semelhança do trabalho descrito por Thompson *et al*.[17], mas o modelo utilizado não inclui nenhum parâmetro associado à probabilidade de associação. No entanto, postula-se que modelação do encolhimento é uma forma de alterar a probabilidade de associação de um SNP visto que diferentes graus de encolhimento fazem variar a probabilidade de um SNP ser significativamente diferente de zero.
Os resultados obtidos indicam a vantagem tanto da análise conjunta de SNPs *per se*, como da inclusão de informação biológica externa. A vantagem da análise conjunta de SNPs já foi previamente

demonstradas, pelo que seria de esperar que o modelo utilizado fosse melhorar a deteção de SNPs e blocos de DL. A inclusão de informação biológica foi também estudada, apesar de em menor escalar que a análise conjunta de SNPs, tendo-se observado uma melhoria na deteção de SNPs quando se inclui informação biológica externa numa análise SNP-a-SNP[17]. Este trabalho unifica as duas abordagens e os primeiros resultados ilustram o potencial benefício da utilização de ambas.

A inclusão da estrutura de blocos de DL constitui um aspeto relevante a referir visto, em geral, esta estrutura não ser tida em conta em métodos de análise de GWAS. Recentemente, tem-se verificado um aumento do número de SNPs identificados no genoma (~8 milhões de variantes) graças às novas tecnologias de sequenciação. A identificação de um maior número de SNPs tem a vantagem de aumentar a probabilidade de detetar o SNP "causador" do fenótipo/doença. No entanto, o aumento do número de variáveis limita ainda mais a utilização de melhores métodos estatísticos para análise dos dados. A consideração da estrutura de blocos de DL permite explorar a correlação entre os SNPs para melhorar a estimação. Uma alternativa a considerar, é a utilização dos blocos de DL como unidade de análise, o que permitira reduzir significativamente o número de variáveis. No entanto, este tipo de abordagem dificulta a inclusão de informação biológica externa, visto que não existe muita informação disponível relativamente aos blocos, especificamente, mas sim relativamente a SNPs e a genes.

Um dos aspetos a salientar relativamente ao modelo desenvolvido por Yi *et al*. é a sua eficiência computacional. O modelo foi muito eficaz na análise dos 2614 SNPs utilizados na análise. O modelo 1, no qual não foi considerado o agrupamento em blocos de DL, foi o modelo que demorou mais tempo computacional (~20 minutos), enquanto os restantes modelos demoraram entre 2 a 6 minutos a completar. O modelo 1 demorou significativamente mais tempo visto o processo de estimação não ser efetuado ao nível do grupo, com menos variáveis em cada bloco. A estimação por grupo, requere menos iterações para atingir a convergência porque o número de SNPs em cada cada grupo é menor, permitindo o algoritmo correr mais rapidamente[13]. Estes resultados revelam o potencial para a utilização da estrutura de blocos de DL, uma estrutura com significado biológico, para diminuir o tempo computacional em modelos de análise conjunta de SNPs. Tal como referido previamente, a falta de eficiência computacional tem sido uma das grandes limitações da implementação de novos modelos na análise de GWAS, principalmente dos modelos Bayesianos. Apesar de se esperar que a aplicação deste modelo à escala de um GWAS não seja simples, este modelo apresenta uma performance computacional muito melhor que outros previamente descritos[18] e, como tal, constitui uma hipótese promissora para atingir esse objetivo.

Os resultados obtidos são sugestivos do impacto que a inclusão informação biológica externa pode ter na detecção de SNPs em estudos de associação genética. No entanto, é ainda necessário desenvolver trabalho relativamente ao tipos de informação que deve ser incluída, à importância relativa a atribuir a cada item de informação e às melhores formas de incluir a informação em modelos Bayesianos, quer seja através da modelação do encolhimento, da modelação dos efeitos dos SNPs ou da probabilidade de associação de um SNP com a doença/fenótipo.

5. Referências

1. Lewis CM, Knight J. Introduction to genetic association studies. *Cold Spring Harb Protoc*. 2012;2012(3):297–306. doi:10.1101/pdb.top068163.
2. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747–53. doi:10.1038/nature08494.
3. De Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. *Nat Genet*. 2005;37(11):1217–23. doi:10.1038/ng1669.
4. Johnstone IM, Titterington DM. Statistical challenges of high-dimensional data. *Philos Trans A Math Phys Eng Sci*. 2009;367(1906):4237–53. doi:10.1098/rsta.2009.0159.

5. Dasgupta A, Sun Y V, König IR, Bailey-Wilson JE, Malley JD. Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. *Genet Epidemiol*. 2011;35 Suppl 1:S5–11. doi:10.1002/gepi.20642.
6. Waldmann P, Mészáros G, Gredler B, Fuerst C, Sölkner J. Evaluation of the lasso and the elastic net in genome-wide association studies. *Front Genet*. 2013;4:270. doi:10.3389/fgene.2013.00270.
7. Gabriel SB. The Structure of Haplotype Blocks in the Human Genome. *Science (80-)*. 2002;296(5576):2225–2229. doi:10.1126/science.1069424.
8. Trégouët D-A, König IR, Erdmann J, et al. Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nat Genet*. 2009;41(3):283–5. doi:10.1038/ng.314.
9. Ehret GB, Lamparter D, Hoggart CJ, Whittaker JC, Beckmann JS, Kutalik Z. A multi-SNP locus-association method reveals a substantial fraction of the missing heritability. *Am J Hum Genet*. 2012;91(5):863–71. doi:10.1016/j.ajhg.2012.09.013.
10. Paré G, Asma S, Deng WQ. Contribution of Large Region Joint Associations to Complex Traits Genetics. *PLOS Genet*. 2015;11(4):e1005103. doi:10.1371/journal.pgen.1005103.
11. Vannucci M, Stingo F. Bayesian Models for Variable Selection that Incorporate Biological Information. *Bayesian Stat*. 2010;9:1–20. Available at: http://www.stat.rice.edu/~marina/papers/VannucciValencia.pdf.
12. Minelli C, De Grandi A, Weichenberger CX, et al. Importance of Different Types of Prior Knowledge in Selecting Genome-Wide Findings for Follow-Up. *Genet Epidemiol*. 2013;37(2):205–213. Available at: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=emed11&AN=2013043270.
13. Yi N, Liu N, Zhi D, Li J. Hierarchical generalized linear models for multiple groups of rare and common variants: Jointly estimating group and Individual-Variant effects. *PLoS Genet*. 2011;7(12). doi:10.1371/journal.pgen.1002382.
14. Burney PG, Luczynska C, Chinn S, Jarvis D. The European Community Respiratory Health Survey. *Eur Respir J*. 1994;7(5):954–60. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8050554. Accessed December 4, 2014.
15. Locke AE, Kahali B, Berndt SI, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*. 2015;518(7538):197–206. doi:10.1038/nature14177.
16. Taliun D, Gamper J, Pattaro C. Efficient haplotype block recognition of very long and dense genetic sequences. *BMC Bioinformatics*. 2014;15(1):10. doi:10.1186/1471-2105-15-10.
17. Thompson JR, Gögele M, Weichenberger CX, et al. SNP prioritization using a Bayesian probability of association. *Genet Epidemiol*. 2013;37(2):214–21. doi:10.1002/gepi.21704.

Leis que governam a estrutura primária do ADN dos seres vivos

Vera Afreixo[1, 2], *vera@ua.pt*
Ana Helena MP Tavares[1], *ahtavares@ua.pt*

1. *Universidade de Aveiro*
2. *iBiMED (instituto de Biomedicina)*

Introdução

Qualquer computador pessoal é capaz de gerar sequências de As, Cs, Gs e Ts com tamanhos iguais aos dos genomas dos seres vivos e até com palavras genómicas com probabilidades iguais às dos genomas reais. No entanto, até um supercomputador, a menos da produção de uma cópia, é incapaz de produzir genomas funcionais. Atualmente ainda se está longe de encontrar modelos capazes de gerar sequências de nucleótidos que sejam funcionais!

O que distancia uma sequência aleatória de nucleótidos de um genoma real? Se assumirmos que os genomas foram criados a partir de uma junção aleatória de nucleótidos esta divergência entre o genoma real e o aleatório poderá ter informação sobre o processo evolutivo das espécies. A descoberta do modelo gerador do genoma de cada espécie poderia trazer grandes contribuições para a caracterização do processo evolutivo das espécies. Mas este é, aparentemente, um problema muito complexo.

É prática comum tentar encontrar a solução para um problema complexo através da decomposição em sub-problemas mais simples.

Neste texto, essencialmente, abordaremos o sub-problema da caracterização das distâncias entre símbolos genómicos e o das extensões da segunda lei de Chargaff avalia a divergência de comportamentos das sequências reais e aleatórias.

Uma sequência simples de ADN pode ser vista como uma sequência de quatro letras A, C, G, T onde A-T e C-G são pares de bases complementares. Uma palavra genómica é um subconjunto de letras que surgem justapostas na sequência de ADN em que o tamanho da palavra é dado pelo número de letras que a compõe.

O complemento invertido de uma palavra genómica é uma palavra que se obtém da palavra inicial alterando cada um dos nucleótidos pelo nucleótido complementar e invertendo a ordem pela qual os nucleótidos se apresentam (e.g. o complemento invertido da palavra ACCGT é ACGGT).

Numa sequência genómica a distância entre nucleótidos pode ser definida como a diferença entre as posições de duas ocorrências sucessivas do mesmo símbolo. Para concretizar esta ideia, considere-se a sequência AACGTCGAAATCCGTAA cujas quatro sequências de distâncias entre nucleótidos são $d^A = (1; 6; 1; 1; 6; 1)$; $d^C = (3; 6; 1)$; $d^G = (3; 7)$; $d^T = (6; 4)$. De modo análogo pode determinar-se a distância entre palavras genómicas (Afreixo *et al.*, 2015a).

O afastamento de uma sequência genómica real a uma sequência gerada num cenário aleatório pode traduzir, de algum modo, a evolução natural das espécies defendida por Darwin. O efeito da aleatoriedade pode ser avaliado através da dissimilaridade entre a sequência real e a sequência gerada num contexto aleatório controlado, em particular, gerada sobre o pressuposto da independência entre símbolos.

Neste artigo será discutida a evolução seletiva dos seres vivos na estrutura primária dos seus genomas em dois aspetos que são usados para caracterizar os genomas das espécies: a relação entre as ocorrência das palavras genómicas complementos invertidos entre si (extensões da segunda lei de Chargaff) e a distribuição de distâncias entre palavras genómicas.

Extensões da segunda lei de Chargaff

As extensões da segunda lei de Chargaff têm vindo a ser referidas como uma lei universal que está presente nos genomas das espécies, sendo este fenómeno também designado por simetria em cadeia simples de ADN (designação adotada). Esta simetria é caracterizada pela semelhança entre as frequências das palavras genómicas e dos correspondentes complementos invertidos. A Fig. 1 apresenta um exemplo relativo à sequência completa do genoma humano para as palavras de tamanho três (também designadas por trinucleótidos). Na figura observa-se que os pares de palavras que são complementos invertidos entre si apresentam frequências de ocorrência próximas e que, de modo geral, apresentam maiores diferenças em relação às frequências de ocorrência das restantes palavras do genoma.

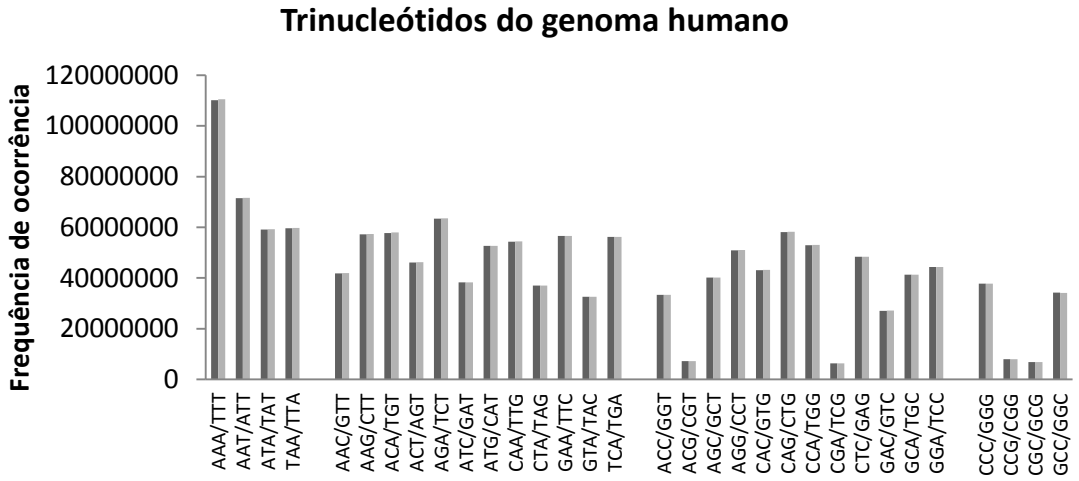


Figura 1: Frequência de ocorrência dos trinucleótidos no genoma humano (versão 37.3).

A avaliação desta regra tem sido feita usando muitas abordagens diferentes, desde a simples correlação entre as frequências das palavras e dos seus complementos invertidos (Baisnée *et al.*, 2002) até abordagens mais exigentes do ponto de vista computacional como a distância de simetria entre palavras baseada na medida de Ulam (Aldous and Diaconis, 1999).
A Tabela 1 apresenta dados relativos à avaliação da simetria em cadeia simples de ADN do genoma humano, para as palavras genómicas até tamanho 10, usando o coeficiente de Pearson, $r \in [-1,1]$, e uma distância baseada na medida de Ulam proposta em Afreixo et al. (2013), $W_s \in [0,1]$.

k	1	2	3	4	5	6	7	8	9	10
r	1	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999
Ws	0	0	0	0,1563	0,3516	0,6763	0,9221	0,9795	0,9967	0,9983
VR	1	210,8	247,4	262,9	270	255,6	214,1	165,3	126,8	103,5

Tabela 1: Avaliação da simetria em cadeia simples de ADN do genoma humano, para as palavras genómicas até tamanho 10, usando o coeficiente de correlação de Pearson (*r*), uma distância baseada na medida de Ulam (W_s) e uma medida de simetria excecional (*VR*).

Os resultados da correlação mostram efeitos de simetria muito fortes, mas estes valores são de certa forma enganadores uma vez que as frequências de ocorrência das palavras constituídas só por As e só por Ts apresentam ordem de grandeza diferente das restantes palavras[1]. A distância W_s é mais robusta

[1] Nas sequências genómicas ocorrem frequentemente poli-As (subsequências de sucessivos As) e poli-Ts (subsequências de sucessivos Ts).

a esta diferença da frequência entre palavras e mostra que para palavras até tamanho quatro o efeito de simetria é forte, mas a partir daí diminui substancialmente. A distância W_s também não é uma medida perfeita de simetria, pois a presença de pequenas variações entre as frequências das diferentes palavras pode contribuir para que esta assuma valores elevados.

Acredita-se que a ocorrência deste tipo de simetria tem motivação biológica como, por exemplo, as estruturas de "stem-loop". Porém, a prevalência de um só fenómeno biológico, por si só, poderá ser insuficiente para explicar a ocorrência do fenómeno de simetria em cadeia simples de ADN. E por outro lado, um modelo aleatório simples poderá gerar sequências que evidenciam um forte efeito de simetria em cadeia simples de ADN, e.g. modelo de independência entre nucleótidos assumindo probabilidades de ocorrência iguais para nucleótidos complementares. Neste caso, constata-se que os grupos de composição equivalente[2] são constituídos por palavras de igual probabilidade e, em particular, as palavras que são complemento invertido entre si têm a mesma probabilidade.
Motivado pela afirmação de Qi *et al.* (2004) que reconhece que a divergência entre o genoma real e o aleatório traduz a evolução seletiva, foi proposta por Afreixo *et al.* (2015) a análise da simetria excecional realçando a semelhança das frequências das palavras genómicas complemento invertidos entre si, com a frequência das palavras genómicas de composição equivalente. A medida de simetria excecional também incorpora conceitos relacionados com testes e medidas de tamanho do efeito de ajustamento, confrontando distribuições empírica e de referência.
A medida de simetria excecional, *VR*, assume valores positivos, o valor 1 não traduz qualquer simetria excecional e valores maiores do que 1 traduzem a simetria excecional (Afreixo *et al.*, 2015b). A Tab.1 apresenta os resultados da medida de simetria excecional, *VR*. Naturalmente, verifica-se que para os nucleótidos não há simetria excecional, enquanto que para as restantes palavras até tamanho 10 existe.
Com esta medida não é possível identificar a origem do fenómeno de simetria em cadeia simples de ADN, mas introduz-se mais conhecimento sobre o fenómeno avaliando, simultaneamente, a semelhança entre as frequências das palavras complemento invertido entre si e a divergência do fenómeno de simetria em estrutura de independência entre nucleótidos. Desta forma através da medida de simetria excecional, acreditamos que estamos a medir a evolução seletiva das palavras no sentido de favorecer mais ou menos o fenómeno da simetria em cadeia simples de ADN.,

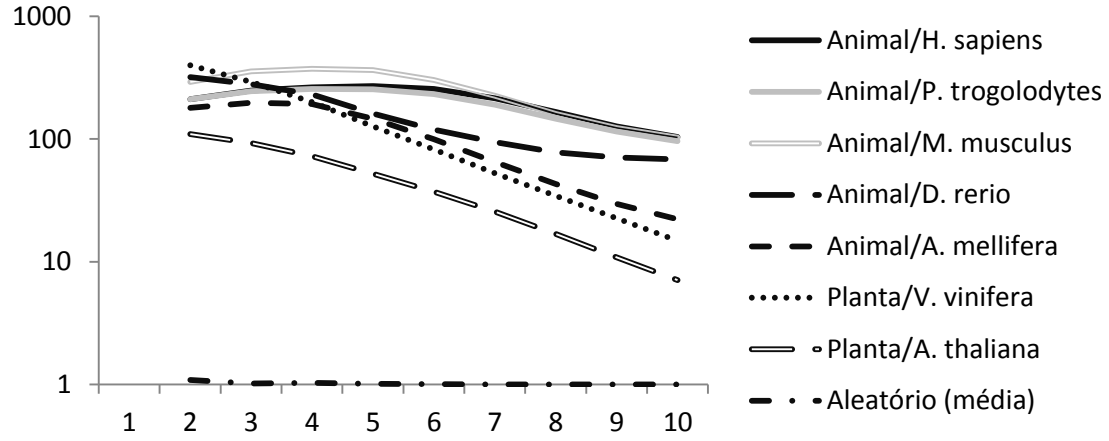


Figura 2: Gráficos de linhas dos valores de simetria excecional para palavras até tamanho 10 (para palavras de tamanho um VR=1). Estão representados seis animais, duas plantas e a média de 100 sequências geradas em contexto de independência, onde as probabilidades dos nucleótidos foram estimadas usando as frequências do genoma humano.

[2] Um grupo de composição equivalente é um conjunto de palavras genómicas que na sua composição têm o mesmo número de A ou T e o mesmo número de C ou G (e.g. as palavras AAC,TCA e TTG pertencem a um mesmo grupo de composição equivalente).

A Fig. 2 apresenta os valores de simetria excecional para diferentes espécies (seis animais incluindo o homo sapiens e duas plantas) e a média dos valores da medida de simetria excecional de 100 sequências aleatórias geradas em contexto de independência de nucleótidos onde a probabilidade de cada nucleótido é estimada pela probabilidade de ocorrência no genoma humano.
Observa-se que há uma clara diferença entre os genomas reais e os aleatórios, sendo que os genomas reais apresentam simetria excecional. Os valores de simetria excecional variam com o tamanho da palavra genómica e apresentam diferenças entre as espécies, de destacar que os gráficos de linhas mais parecidos são os do genoma humano (H. sapiens) e do genoma do chimpanzé (P. trogolodytes).

Distâncias entre palavras genómicas

As distribuições de distâncias entre palavras genómicas têm vindo a ser exploradas em diferentes contextos, tais como a comparação do genoma de diferentes espécies, bem como a distinção entre as diferentes partes constituintes do ADN (e.g. ADN mitocondrial, ilhas de CpG).
A partir da sequência de distâncias, para uma determinada palavra, é então possível definir a distribuição empírica das distâncias entre palavras. No exemplo AACGTCGAAATCCGTAA, e para o caso das palavras AA, as distâncias 1 e 7 têm uma frequência relativa de 1/3 e 2/3, respetivamente.
Uma forma de determinar propriedades estatísticas de diferentes genomas é estudar e caracterizar a distribuição de distâncias entre palavras genómicas. Assumindo que a sequência de palavras é gerada aleatoriamente e que as palavras são independentes e identicamente distribuídas, a distribuição de distâncias entre palavras é a distribuição geométrica.
No entanto, na maioria dos estudos as palavras genómicas são contadas tendo em conta a sobreposição de palavras adjacentes, sendo óbvia a não independência entre as palavras da sequência. Assim, numa abordagem que assuma a sobreposição de palavras, impõe-se a definição de outras distribuições de referência.
A distância entre ocorrências sucessivas de padrões de letras é um assunto estudado, com muitos resultados teóricos já deduzidos, em particular, no que diz respeito à função que traduz o tempo de espera até ao retorno a um padrão específico (e.g. Robin *et al.* (2001)). No caso mais simples em que se pretende obter a distribuição de distâncias entre palavras em sequências cujos nucleótidos são gerados independentemente, e admitindo a sobreposição entre palavras, poderemos recorrer a um diagrama de estados que traduz o progresso na identificação da palavra à medida que cada símbolo é lido na sequência (ver Afreixo *et al.* (2015a)). Sob o pressuposto de independência dos nucleótidos, qualquer transição para um estado está associada à probabilidade de ocorrência do novo nucleótido que é lido na sequência. Note que cada distância está associada ao número de transições necessárias até obter o sucesso (o estado absorvente) e aos diferentes percursos que se podem efetuar até o obter.
Numa sequência genómica real, as distribuições de distâncias entre palavras são diferentes das distribuições associadas a uma sequência gerada aleatoriamente em contexto de independência entre os nucleótidos. Por exemplo, no humano, verifica-se que a distribuição de distâncias associada a uma palavra é muito semelhante à distribuição de distâncias associada ao seu complemento invertido (Tavares *et al.*, 2015) e que a distribuição de distâncias do CG apresenta características muito diferentes das distribuições dos restantes nucleótidos (Afreixo *et al.*, 2015).
A Fig. 3 apresenta as distribuições de distâncias da palavra CG do genoma humano e da distribuição de referência em contexto de independência entre nucleótidos, evidenciando-se algumas diferenças: até cerca da distância 50 as frequências empíricas são inferiores às do modelo de referência e para distâncias maiores são superiores.
A geração de modelos para a distribuição de distâncias entre palavras em contexto aleatório apresenta-se como tópico de estudo de interesse, e acredita-se que permite evidenciar a contribuição da evolução seletiva no ADN das espécies.
De modo a avaliar a divergência de comportamentos entre sequências reais e sequências geradas em cenários aleatórios, deve subtrair-se o efeito da aleatoriedade à sequência real (Qi *et al.*, 2004). A análise de resíduos apresenta algum potencial na construção de árvores filogenéticas. Por exemplo, na comparação entre a distribuição de distâncias real e a distribuição de referência pode recorrer-se ao erro relativo (Afreixo *et al.*, 2009).

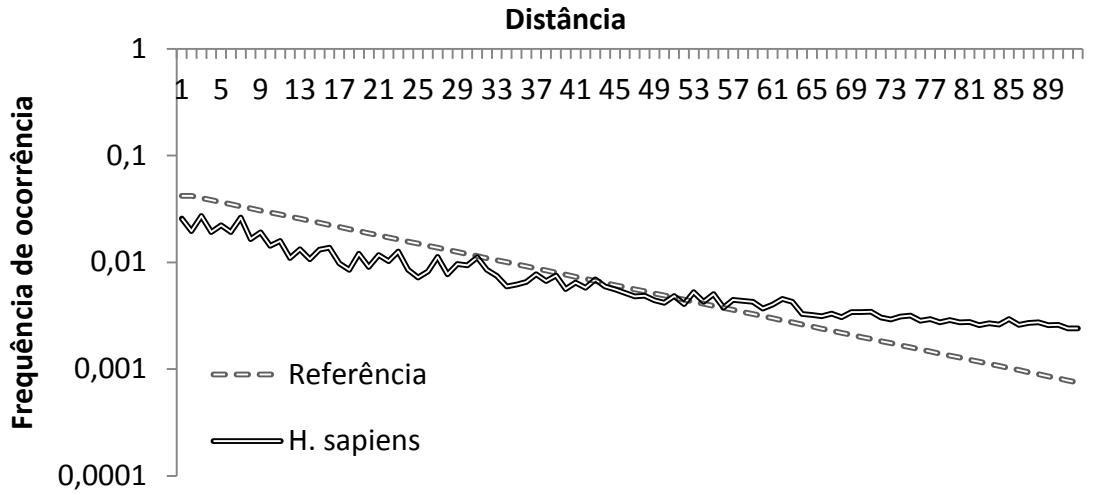


Figura 3: Gráficos de linhas das distribuições de distâncias da palavra CG do genoma humano e da distribuição de referência em contexto de independência entre nucleótidos.

O vetor de erros relativos associado a um genoma pode ser visto como uma assinatura genómica que identifica a espécie, podendo ser usados na construção de dendrogramas muitas vezes interpretados como árvores filogenéticas. As espécies são hierarquicamente agrupadas e as dissimilaridades podem ser vistas como distâncias evolutivas. Por exemplo, se o objetivo for comparar as espécies tendo por base as distribuições de distâncias entre palavras de um determinado tamanho, k, pode definir-se uma distribuição global, bastando para tal aplicar a lei de probabilidade total a todas as 4^k distribuições. O dendrograma da Fig. 4 foi construído com base nas distribuições de distâncias globais entre dinucleótidos, para genomas completos, aplicando o método de ligação média e usando a distância euclidiana no cálculo da matriz de similaridade. O corte em quatro classes evidencia uma separação das espécies em quatro grupos bem distintos: animais, plantas, fungos e bactérias.

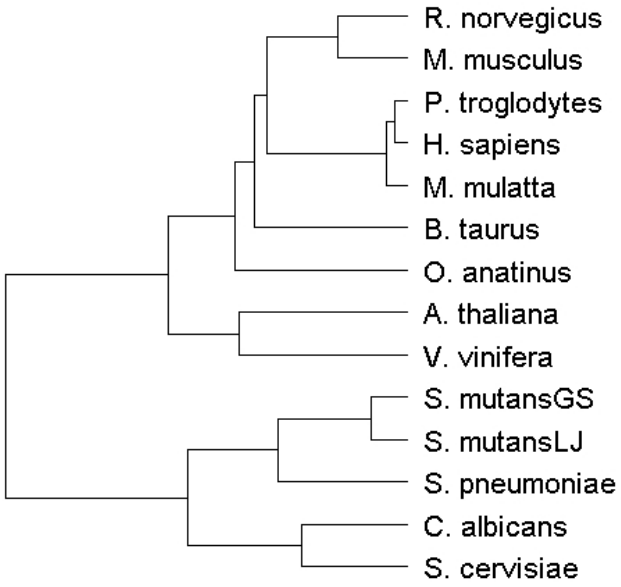


Figura 4: Dendrograma das assinaturas genómicas (definidas a partir dos erros relativos entre as distribuições das distâncias globais das palavras de tamanho 2 e a distribuição de referência global obtida em contexto de independência de nucleótidos). Foram considerados sete animais (R. norvegicus, M. musculus, P. troglodytes, H. sapiens, M. mulatta, B. taurus, O. anatinus), duas plantas (A. thaliana, V. vinífera), dois fungos (C. albicans, S. cervisiae) e três bactérias (S. mutansGS, S. mutansLJ, S. pneumoniae).

Conclusão

A partir dos conceitos de distâncias entre palavras e de simetria em cadeia simples de ADN é possível extrair perfis, diretamente relacionados com as características do ADN, mostrando potencial na discriminação entre espécies e na caracterização de estrutura primária do ADN das espécies.

A explicação da evolução seletiva dos organismos vivos é um tópico que provavelmente não terá uma resposta única e naturalmente não há forma de avaliar a veracidade das conclusões que tenham e venham a ser tiradas. No entanto, poder avaliar o desempenho de um procedimento na análise de dados genómicos tem sido uma experiência fantástica para os investigadores que abraçam estes temas. A criação de assinaturas genómicas poderá ter grande potencial na identificação automática de espécies, principalmente em áreas tão importantes como a da microbiologia.

Referências

Afreixo, V., Bastos, C. A. C., Garcia, S. P., Rodrigues, J. M. O. S., Pinho, A. J., Ferreira, P. J. S. G., 2013. The breakdown of the word symmetry in the human genome. *Journal of Theoretical Biology* 335, 153–159.

Afreixo, V., Bastos, C. A. C., Rodrigues, J. M. O. S., Silva, R. M. 2015a. Identification of DNA CpG islands using inter-dinucleotide distances. *Optimization in the Natural Sciences: Communications in Computer and Information Science* 499, 162–172.

Afreixo, V., Rodrigues, J.M., Bastos, C.A.C., 2015b. Analysis of single-strand exceptional word symmetry in the human genome: new measures. *Biostatistics*, 16, 209–221.

Afreixo, V., Bastos, C. A. C., Pinho, A. J, Garcia, S. P. and Ferreira, Paulo J. S. G., 2009. Genome analysis with inter-nucleotide distances. *Bioinformatics*, 23, 3064–3070.

Aldous, D., Diaconis, P., 1999. Longest increasing subsequences: from patience sorting to the Baik–Deift–Johansson theorem. *Bull. Am. Math. Soc*. 36, 199–213.

Baisnée, P.F., Hampson, S., Baldi, P., 2002. Why are complementary DNA strands symmetric? *Bioinformatics* 18, 1021–1033.

Qi, J., Wang, B., Hao, B.I., 2004. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *Journal of Molecular Evolution* 58, 1–11.

Robin, S., Daudin J.J., 2001. Exact distribution of the distance between any occurrences of a set of words. *Annals of the Institute of Statistical Mathematics*, 53 (4), 895–905.

Tavares, A.H., Afreixo, V., Rodrigues, J. M. O. S., Bastos, C. A. C., 2015. The symmetry of oligonucleotide distance distributions in the human genome. *ICPRAM*, Vol. 2 pp. 256-263, Science and Technology Publications.

Pós-Doc

Tese de Doutoramento: Modelação e Avaliação de Desempenho de Redes Móveis Ad Hoc

(Boletim SPE outono de 2011, p. 95)

Gonçalo Jacinto, *gjcj@uevora.pt*

Universidade de Évora, Escola de Ciências e Tecnologia, Departamento de Matemática e CIMA-UE

Parece que foi ontem! Na minha memória estão a maior parte dos detalhes do processo que culminou com as provas de Doutoramento, mas a realidade é que já passaram 4 anos e exatamente sete meses (à data que escrevo este texto). Realmente diziam que era um processo muito duro. Verdade seja dita, foi! Não tanto nos primeiros tempos em que apenas me preocupava em obter novos resultados e a fazer investigação pura, mas sim na parte final, com a escrita da tese, a revisão final da tese, a espera pela parte burocrática e a ansiedade da apresentação das provas.

Realizei as provas no último dia de Fevereiro de 2011 no Instituto Superior Técnico, hoje pertencente à Universidade de Lisboa. Tenho que aproveitar este texto para agradecer publicamente aos meus orientadores, os Professores António Pacheco e Nelson Antunes, não só pela oportunidade que me deram para poder aprender com eles, mas também por me terem ensinado todo o processo de rigor científico, de terem contribuído para o meu desenvolvimento pessoal e científico e incentivarem a melhoria contínua. A eles, cujo trabalho e amizade já vem desde a tese de Mestrado e continua, o meu muito Obrigado!

Olhando para trás, muito para trás, e sendo totalmente honesto, nunca me imaginei a terminar um Doutoramento em Matemática e muito menos vir a ensinar Probabilidades e Estatística numa Universidade. Hoje, ao contrário de mim que não imaginava tal cenário porque simplesmente nunca o coloquei na minha adolescência, é triste verificarmos que existem bastantes adolescentes que não imaginam tal cenário porque simplesmente não existem oportunidades para tal. Em termos futuros, este cenário levará a uma grande machadada na capacidade de Portugal construir mais uma geração de estatísticos e de continuar o bom trabalho desenvolvido nas gerações anteriores.

Bem, deixemos os lamentos para trás. Sinceramente, a conclusão do Doutoramento pouco alterou a minha vida profissional e pessoal. Para isso contribuiu o facto de já estar integrado na carreira docente Universitária no início do Doutoramento, como docente no Departamento de Matemática da Universidade de Évora. Aumentei um pouco as minhas responsabilidades com cientista e docente, mas pouco mais. Também o facto de nunca ter gostado de tratar as pessoas pelo seu título, gostando de tratar colegas de trabalho e até alunos por "tu" (embora ainda tenha dificuldade em fazê-lo com os mais veneráveis professores com quem me cruzei e tive o prazer de trabalhar e conviver, como são os exemplos, entre outros, dos Professores Carlos Braumann e António Pacheco – talvez um dia), acabou por contribuir para que a conclusão do Doutoramento não fizesse qualquer diferença na minha vida pessoal e profissional.

Na tese de doutoramento dediquei-me à modelação e análise de desempenho de redes móveis ad hoc. As redes móveis ad hoc são constituídas por nós que se organizam de forma autónoma e sem qualquer infraestrutura, onde a mobilidade dos nós e a possibilidade de comunicação por rotas com múltiplos passos torna a topologia destas redes dinâmica e imprevisível. Têm inúmeras aplicações, desde a constituição de uma rede de telecomunicações em zonas que foram atingidas por catástrofes (onde as redes de telecomunicações existentes são totalmente destruídas), passando pela monitorização do meio ambiente e de animais selvagens, até às redes de veículos, onde estes comunicam entre si para fornecer informações de segurança aos condutores. A parte mais importante da minha tese de doutoramento foi o desenvolvimento de um processo de Markov determinístico por troços que pudesse caraterizar a dinâmica das rotas de comunicação. Foram derivadas a distribuição e o tempo médio de duração das rotas através de um sistema de equações integro-diferenciais. Os resultados obtidos foram muito interessantes, e muito mais resultados ficaram por concluir. Mas talvez o extenuante trabalho computacional que foi desenvolvido tivesse desencorajado, para já, a voltar a olhar para este modelo. Talvez um dia!

Além disso, resultados sobre a conectividade em redes unidimensionais (para as aplicações para as redes de veículos) e bidimensionais (para as aplicações para a monitorização de populações animais e do meio ambiente) foram também obtidos na tese de Doutoramento. Apenas o problema de conectividade no plano, que ficou numa fase mais embrionária, teve continuidade no trabalho pós-doc. Conseguimos obter novos resultados e construir um modelo totalmente novo, que permite obter a probabilidade de dois móveis, de um conjunto de móveis distribuídos no plano sobre um processo de Poisson, estejam conectados através de um número determinado de passos.

A par deste trabalho tenho continuado a trabalhar com os meus ex-orientadores no âmbito da modelação e análise de desempenho na área das telecomunicações. Temos vindo a trabalhar na estimação das características do tráfego da Internet usando a informação parcial que se obtém de um router através de pacotes especiais de tráfego (chamados de sondas), que permite estimar as características do tráfego real através do desempenho observado pelas sondas. No seguimento do mesmo, estamos neste momento a derivar novos resultados para estimar o tráfego numa rede de routers em vez de apenas num router e a estudar a estimação da taxa de utilização de cada router. Além disso, estamos também a investigar a possibilidade de derivar a distribuição de probabilidade do tempo de serviço de um router através de técnicas de decomposição, que consistem na inversão de um processo de Poisson composto ou através da reversão de uma série de potências.

Além do trabalho que tenho feito na área em que me doutorei, por força das orientações de alunos de Mestrado do Mestrado em Modelação Estatística e Análise de Dados que ministramos na Universidade de Évora, e pelo esforço que fazemos para que os melhores alunos submetam comunicações aos congressos científicos nacionais, acabei por começar a trabalhar nos últimos anos na modelação estatística e análise de dados. Quer dizer, na verdade não foram os alunos os catalisadores, mas sim o meu grande amigo e colega de gabinete Paulo Infante, que me tem cativado e convidado para participar nas inúmeras parcerias que tem desenvolvido com as instituições locais. Tem sido através da colaboração científica com ele e a Anabela Afonso, que desenvolvemos protocolos de colaboração com a Câmara Municipal de Évora e com o Agrupamento de Centros de Saúde do Alentejo Central (ACES), nos quais avaliámos a eficácia do programa Seniores Ativos, um programa de atividade física oferecido pela Câmara Municipal de Évora aos indivíduos com 55 ou mais anos de idade, para que os seniores beneficiem de melhor qualidade de vida através da atividade física regular, quer a nível social quer a nível de saúde. Também através do protocolo estabelecido com o ACES, foi estabelecida a colaboração no âmbito do projeto AdolesSer, onde foi avaliada a eficácia das ações de formação para a saúde sexual e reprodutiva nas escolas de Évora para os jovens a partir do 6º ano de escolaridade. Estamos ainda a trabalhar noutros projetos, mas ainda estão numa fase inicial, pelo que não valerá a pena falar deles aqui.

Através de um amigo que tive o prazer de conhecer enquanto colegas de gabinete durante o nosso doutoramento, conheci uma pessoa que necessitava de apoio estatístico para a sua tese de Mestrado na área da neuropsicologia. Dessa colaboração têm resultado diversos trabalhos científicos nos quais

desenvolvo a componente de modelação estatística. Tem sido uma experiência verdadeiramente interessante, até porque a Daniela e o Humberto, encaixam que nem uma luva no lamento que efetuei no início deste texto. Eles foram obrigados a emigrar para a Colômbia para poderem fazer aquilo que mais gostam de fazer: investigação científica, um em matemática e outro em neuropsicologia. Tem sido uma colaboração bastante interessante, pois permitiu a colaboração entre a Universidade de Évora, a Universidade El Bosque, na Colômbia, e a Universidade Autónoma de Barcelona, apesar das restrições orçamentais com que nos deparamos, que embora sendo inviável no passado, no presente se torna possível devido às novas tecnologias.

Ao nível das tarefas de gestão Universitária, para além de cargos que tenho ocupado nas Comissões de Curso e outros de menor importância, têm sido as atividades de divulgação Estatística e Matemática que temos desenvolvido em Évora, que me têm dado mais prazer em desenvolver. Com estas ações pretendemos cativar o interesse dos alunos do Secundário pela área da matemática e, felizmente, parece que estamos a ver uma luz ao fundo do túnel porque este ano conseguimos uns incríveis 34 alunos no curso de matemática em Évora, algo que já não acontecia desde …. o início do milénio!

Assim se passaram quatro anos e exatamente sete meses após o meu doutoramento. Nos últimos 6 meses com uma nova alegria e novas tarefas ainda mais exigentes. Fui Pai pela primeira vez! Veremos se daqui a 18 anos esta nuvem negra que paira sobre o panorama científico português já passou e o meu filho já poderá escolher livremente a área onde quer estudar e trabalhar, sem que com isso seja obrigado a emigrar!

Gonçalo Jacinto

Ciência Estatística

• Livros

Título: *Programa e Resumos. XXII Congresso SPE 2015*
Editores: Clara Cordeiro, Conceição Ribeiro, Maria Helena Gonçalves e Nelson Antunes
Ano: 2015. Editora: Sociedade Portuguesa de Estatística. ISBN: 978-972-8890-36-0

Título: *Estatística Bayesiana Computacional – uma introdução*
Autores: M. Antónia Amaral Turkman e Carlos Daniel Paulino
Ano: 2015. Editora: Sociedade Portuguesa de Estatística. ISBN: 978-972-8890-37-7

• Teses de Doutoramento

Título: Clustering with Discrete Mixture Models - An integrated approach for model selection
Autora: Cláudia Marisa Vasconcelos Silvestre, *csilvestre@escs.ipl.pt*
Orientadores: Mário Alexandre Teles de Figueiredo e Margarida G. M. S. Cardoso

A minha tese enquadra-se na área da Análise de Agrupamento de dados categorizados. A Análise de Agrupamento tem sido amplamente usada em inúmeras áreas de aplicação e, como consequência, têm surgido continuamente novos desafios que motivam a produção científica nesta área. Grande parte do trabalho sobre esta temática limita-se à análise de dados contínuos. No entanto, nomeadamente em ciências sociais, é frequente ter que lidar com dados categorizados colocando-se novos desafios à sua análise.

A tese tem dois objetivos principais: 1) identificar o número de grupos subjacente a dados categorizados e 2) identificar um subconjunto das variáveis mais relevantes para o agrupamento dos mesmos dados. Para o agrupamento é usada a estimação de modelos de mistura finita, um enquadramento que torna possível a abordagem formal dos dois objetivos anteriormente referidos, pontos esses que permanecem como objeto de interesse. Neste contexto pressupomos que os dados (categorizados) provêm de uma mistura de distribuições multinomiais e usamos o critério *Minimum Message Length* (MML) para selecionar um modelo parcimonioso que descreva bem os dados.

Na nossa proposta integramos a determinação do número de grupos, assim com a seleção de variáveis base de agrupamento, no processo de estimação do modelo de mistura finita. Esta abordagem contraria a maioria das abordagens neste domínio em que, tipicamente, a escolha do número de grupos, por exemplo, é feita *a posteriori*. Assim, critérios bem conhecidos como o *Bayesian Information Criterion* (BIC), o *Akaike Information Criterion* (AIC) ou o *Integrated Completed Likelihood* (ICL) necessitam que o agrupamento seja feito previamente, mediante a maximização da função de verosimilhança associada ao modelo de mistura considerado. Numa etapa seguinte, o agrupamento é feito para diferentes números de grupos e escolhe-se a solução que corresponde ao melhor valor do critério pré-estabelecido.

Para a seleção do número de grupos implementamos uma nova variante do algoritmo *Expectation Maximization* (EM) – o EM-MML – que integra, simultaneamente, num único algoritmo, a estimação e a selecção de um modelo de mistura discreto.

Com o objetivo de identificar um subconjunto de variáveis mais relevantes para o agrupamento e pressupondo que o número de componentes da mistura (grupos) é conhecido, desenvolvemos uma outra variante do algoritmo EM. Recorrendo também ao critério MML e, ainda, ao conceito de saliência duma variável, integramos, na variante proposta – EM-Embedded - a estimação dos parâmetros do modelo de mistura e a seleção das variáveis categorizadas para o agrupamento.

Cada um dos procedimentos propostos (EM-MML e EM-Embedded) é testado em dados simulados e dados reais e comparado com outros métodos, revelando um bom desempenho no agrupamento de dados categorizados.

Cláudia Silvestre

Título: Estudo das atitudes em relação à Estatística dos professores do 1º ciclo e dos professores de Matemática do 2º ciclo do ensino básico

Autor: José Alexandre dos Santos Vaz Martins, *jasvm@ipg.pt*
Orientadoras: Assumpta Estrada Roca e Maria Manuel da Silva Nascimento

A minha tese centra-se na medição e caracterização das atitudes em relação à Estatística dos docentes do 1º e do 2º ciclo do ensino básico português usando uma escala já testada e com boas características psicométricas, *Escala de Actitudes hacia la Estadística de Estrada* – EAEE – (Estrada, 2002). Esta temática surge uma vez que para se atingir o sucesso na educação estatística, e em especial num período de mudanças no ensino básico, reconhece-se a necessidade de conhecer as atitudes dos professores em relação à Estatística. Deste modo podem aumentar-se as possibilidades de criar predisposição, vontade e comprometimento dos professores para com as mudanças necessárias no processo de ensino-aprendizagem e na sua formação.

No primeiro capítulo deste trabalho enquadra-se a temática desta tese, fazendo-se uma análise resumida da história do ensino e da formação de professores em Portugal, com ênfase nos ciclos do ensino abrangidos pelo estudo, e destacando-se o ensino da Matemática e da Estatística.

No segundo capítulo começa por fazer-se uma abordagem geral ao conceito de atitudes. Em seguida, analisam-se e enquadram-se as atitudes em relação à Matemática e à Estatística, mas de forma mais extensa, profunda e específica para a Estatística. Apresentam-se também vários instrumentos de medição dessas atitudes, algumas das suas características e vários dos estudos de aplicação dos referidos instrumentos.

No terceiro capítulo apresenta-se o desenho do estudo e a metodologia utilizada. Além disso, aborda-se a aplicação e a caraterização de uma amostra de 1098 professores do 1º e 2º ciclos do ensino básico de três distritos portugueses da versão em português da escala EAEE, na qual se introduziu a possibilidade de os respondentes apresentarem a justificação da sua classificação em nove dos itens da escala.

No quarto capítulo abordam-se os resultados sobre as atitudes ao nível da pontuação global e das suas componentes, que se verificaram ser positivas. Analisam-se também os itens, bem como a fiabilidade e generalização da escala de atitudes usada que podem ser consideradas elevadas. Apresenta-se ainda a relação entre as componentes das atitudes. Complementa-se a análise com a influência das variáveis do estudo sobre as atitudes, verificando-se existirem diferenças significativas em relação à pontuação global nas variáveis ciclo de ensino, tempo de serviço, área de formação inicial ou especialidade, nível de estudo de Estatística e no nível de ensino de Estatística. Além disso, analisam-se as justificações dadas pelos professores nos nove itens escolhidos, que permitem compreender melhor as atitudes através das motivações para as pontuações naqueles itens.

Na conclusão demonstra-se a coerência global da investigação, a discussão dos resultados no quadro das hipóteses e face a outros estudos, bem como o facto de os objetivos terem sido alcançados. Também são discutidas as implicações dos resultados, nomeadamente para a uma pedagogia das atitudes e para intervenções preventivas e corretivas na formação; avaliadas de forma crítica as limitações da investigação; e efetuadas sugestões e recomendações que abrem perspetivas para investigações futuras.

José Alexandre Martins

Título: Spatio-temporal modelling of environmental data
Autor: Luís Margalho, *lmelo@isec.pt*
Orientadoras: Raquel Menezes Mota Leite e Inês Pereira Silva Cunha Sousa

Na minha tese, enquadrada no âmbito da geoestatística e fundamentalmente motivada por estudos de monitorização ambiental, pretendeu-se verificar a relevância, traduzida em termos da precisão de predições, de incorporar num processo de predição espaço-temporal não somente informação resultante de observações do passado, mas também de variáveis explicativas do processo em observação.

Assim, foi proposto (i) uma extensão de um modelo de predição espaço-temporal já existente, considerando covariáveis geo-referenciadas, e (ii) um novo modelo espaço-temporal, adequado para estudos com poucas observações na dimensão temporal, considerando igualmente covariáveis geo-referenciadas, bem como covariáveis explicativas do comportamento temporal do processo. Neste caso, a função de covariância espaço-temporal foi deduzida por forma a permitir, sob a hipótese de separabilidade, valores diferentes do parâmetro de escala nas componentes espacial e temporal, de forma oposta à interpretação mais tradicional para esta função.

A base de dados utilizada neste estudo era constituída por medidas de deposição de metais pesados em musgos, resultantes de três campanhas de amostragem realizadas a nível nacional em 1992, 1996 e 2002.

A precisão das predições efetuadas foi analisada ao caracterizar padrões espaço-temporais de deposição de metais pesados em Portugal continental.

Inicialmente foi efetuada uma análise exploratória descritiva e uma análise exploratória espacial dos dados, utilizando técnicas bem conhecidas de interpolação espacial. De seguida, foi desenvolvida uma previsão espaço-temporal da concentração de metais pesados para a campanha mais recente, permitindo incorporar variáveis geo-referenciadas explicativas do processo sob observação. Para isso, recorri a um modelo de previsão espaço-temporal existente. Este modelo incide sobre a dimensão espacial do processo através da definição de campos aleatórios para a média, para a escala e para os resíduos, e incorporando a dimensão temporal por meio de campos aleatórios estritamente temporais, que funcionam como correções para a evolução temporal do processo.

Motivado pelo fato de o conjunto de dados em uso ser denso na dimensão espacial, mas escasso em termos temporais, foi proposta uma abordagem *model-based* para dados Gaussianos, e que corresponde a um modelo de correlação saturado na dimensão temporal. O modelo proposto foi deduzido por forma a acomodar não somente covariáveis geo-referenciadas, mas também covariáveis associadas ao comportamento temporal do processo.

No que respeita à precisão dos valores de concentração de metais pesados, a comparação das previsões obtidas por meio de modelos puramente espaciais com as obtidas por modelos espaço-temporais revelou um melhor desempenho por parte destes últimos. É de realçar ainda que, se a comparação for restringida aos dois modelos espaço-temporais, a abordagem *model-based* proporcionou melhores resultados.

Luís Margalho

• Prémios Estatístico Júnior 2015 e Prémio aos Cursos CEF/EFA

No passado dia 3 de outubro realizou-se no Porto a Sessão de Entrega dos **Prémios Estatístico Júnior 2015** e do **Prémio aos Cursos CEF/EFA**. Pela primeira vez, a entrega dos Prémios foi realizada num local público o que possibilitou uma maior visibilidade do acontecimento que, tradicionalmente, se realizava durante os congressos.
A cerimónia teve lugar nas instalações da FNAC no Porto.

SPE – SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Sessão de Entrega dos Prémios Estatístico Júnior 2015

3 de outubro de 2015

FNAC do Porto

ORADORES CONVIDADOS

José Paulo Viana

Divertimentos com Probabilidades (e Estatística)


Irene Oliveira

"Cientista de dados" - uma profissão para o século XXI


ENTREGA DOS PRÉMIOS ESTATÍSTICO JUNIOR 2015

Trabalhos classificados em 1º lugar *ex-áqueo* (Ensino Básico)
"À Procura de perturbações do comportamento alimentar...."
Autores: Leonor Veríssimo C. Lince Duarte
Professora orientadora: Teresa Isabel
Escola Artur Gonçalves, Torres Novas

"Consumo de Drogas e Álcool ao longo da Vida"
Autores: Inês Pereira de Figueiredo, Mariana Rocha Ferreira Dias
Professora orientadora: Catarina Raquel Pedrosa Ferreira M. Silva
Escola EB 2,3 da Maia - Agrupamento Gonçalo Mendes da Maia

"Consumo do Tabaco"
Autores: Joana Costa Ribeiro, Beatriz Calheiros Serpa Pinto Caetano, Bárbara Guimarães Machado
Professor orientador: Bernardino Carneiro de Andrade
Escola EB 2,3 da Maia - Agrupamento Gonçalo Mendes da Maia

Trabalho classificado em 2º lugar (Ensino Básico)
"Ambientador "Jovem Essência""
Autor: Cristiano Godinho da Piedade, Vasco Inácio Franco e Alexandre Carraco Lopes
Professor orientador: Carlos André Pimentel Lameirinhas
Colégio Senhor dos Milagres, Leiria









Trabalhos classificados em 3º lugar (Ensino Básico)
"Hábitos dos Adolescentes face aos Telemóveis"
Autores: Gonçalo Moreira Nunes Ferreira, Rui Pedro Ferraz Pinto
Professor orientador: Bernardino Carneiro de Andrade
Escola EB 2,3 da Maia - Agrupamento Gonçalo Mendes da Maia

Trabalho classificado em 1º lugar (Ensino Secundário)
"Doenças cardiovasculares: O que faz para as prevenir?"
Autores: Andreia Filipa Lopes Sousa, Carolina Cabeleira Felgueiras e Inês da Costa Delgado
Professora orientadora: Maria Alice da Silva Martins
Escola Artur Gonçalves, Torres Novas

Trabalho classificado em 3º lugar (Ensino Secundário)
"Hábitos Tabagistas"
Autor: Mariana Sofia Marques Ribeiro, Carolina Gonçalves Arriaga e Sofia Lopes Marques
Professor orientador: Maria do Carmo Silva Margato Gonçalves
Secundária Dr. Bernardino Machado

Prémio aos Cursos CEF/EFA
Título: ***"A Web no quotidiano"***
Autor: Daniel Christian Correia, Henrique Luís Vilhena Polónio e Bruno Filipe Oliveira Marques
Professor orientador: Vanda Filipa Nobre Ferreira
Instituto de Tecnologias Náuticas, Paço d'Arcos, Oeiras

EXPLORÍSTICA




Prémios Estatístico Júnior 2015

Trabalhos classificados em 1º lugar *ex-aequo* (Ensino Básico)

Título: *"À Procura de pertubações do comportamento alimentar..."*
Autores: Leonor Veríssimo C. Lince Duarte
Professora orientadora: Teresa Isabel
Estabelecimento de Ensino: Escola Artur Gonçalves, Torres Novas

Título: *"Consumo de Drogas e Álcool ao longo da Vida"*
Autores: Inês Pereira de Figueiredo, Mariana Rocha Ferreira Dias
Professora orientadora: Catarina Raquel Pedrosa Ferreira M. Silva
Estabelecimento de Ensino: Escola EB 2,3 da Maia - Agrupamento Gonçalo Mendes da Maia

Título: *"Consumo do Tabaco"*
Autores: Joana Costa Ribeiro, Beatriz Calheiros Serpa Pinto Caetano, Bárbara Guimarães Machado
Professor orientador: Bernardino Carneiro de Andrade
Estabelecimento de Ensino: Escola EB 2,3 da Maia - Agrupamento Gonçalo Mendes da Maia

Trabalho classificado em 2º lugar (Ensino Básico)

Título: *"Ambientador "Jovem Essência""*
Autor: Cristiano Godinho da Piedade, Vasco Inácio Franco e Alexandre Carraco Lopes
Professor orientador: Carlos André Pimentel Lameirinhas
Estabelecimento de Ensino: Colégio Senhor dos Milagres, Leiria

Trabalhos classificados em 3º lugar (Ensino Básico)

Título: *"Hábitos dos Adolescentes face aos Telemóveis"*
Autores: Gonçalo Moreira Nunes Ferreira, Rui Pedro Ferraz Pinto
Professor orientador: Bernardino Carneiro de Andrade
Estabelecimento de Ensino: Escola EB 2,3 da Maia - Agrupamento Gonçalo Mendes da Maia

Trabalho classificado em 1º lugar (Ensino Secundário)

Título: *"Doenças cardiovasculares: O que faz para as prevenir?"*
Autores: Andreia Filipa Lopes Sousa, Carolina Cabeleira Felgueiras e Inês da Costa Delgado
Professora orientadora: Maria Alice da Silva Martins
Estabelecimento de Ensino: Escola Artur Gonçalves, Torres Novas

Trabalho classificado em 3º lugar (Ensino Secundário)

Título: *"Hábitos Tabagistas"*
Autor: Mariana Sofia Marques Ribeiro, Carolina Gonçalves Arriaga e Sofia Lopes Marques
Professora orientador: Maria do Carmo Silva Margato Gonçalves
Estabelecimento de Ensino: Escola Secundária Dr. Bernardino Machado

Prémio aos Cursos CEF/EFA

Título: *"A Web no quotidiano"*
Autor: Daniel Christian Correia, Henrique Luís Vilhena Polónio e Bruno Filipe Oliveira Marques
Professora orientador: Vanda Filipa Nobre Ferreira
Estabelecimento de Ensino: Instituto de Tecnologias Náuticas, Paço d'Arcos, Oeiras

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Prémio SPE 2015

Modelação Conjunta de Dados Longitudinais e de Sobrevivência de Cancro da Mama

Ana Borges, *aib@estgf.ipp.pt*

CIICESI, ESTGF — Instituto politécnico do Porto, DMA-ECUM, Universidade do Minho

O trabalho desenvolvido propõe a utilização de métodos estatísticos no âmbito da bioestatística para estudar o cancro da mama, em pacientes da unidade de Senologia no Hospital de Braga. Com a principal motivação de contribuir para a compreensão da progressão do cancro da mama, na população Portuguesa, usando um modelo estatístico com pressupostos mais complexos do que a análise tradicional.

A análise levada a cabo tem como objetivo primordial desenvolver modelos conjuntos para dados longitudinais (medições repetidas ao longo do tempo de marcadores tumorais) e de sobrevivência (tempo até evento de interesse) de pacientes com cancro de mama, sendo a morte por cancro da mama o evento de interesse. Iniciou-se com uma análise exploratória do conjunto de dados de 540 pacientes, englobando 50 variáveis, recolhido dos registos médicos do Hospital. Realizou-se, posteriormente, uma análise de sobrevivência independente a fim de compreender quais os possíveis fatores de risco para a morte por cancro de mama, para estes pacientes. Seguiu-se uma análise longitudinal independente do marcador tumoral Carcinoma Antigénio 15-3 (CA15-3), para identificação de fatores de risco relacionados com o aumento dos seus valores. Para análise de sobrevivência recorreu-se ao modelo de riscos proporcionais de Cox e ao modelo paramétrico flexível de Royston-Parmar. Modelos lineares generalizados de efeitos mistos foram utilizados para estudar a progressão longitudinal do marcador tumoral. Após análises de sobrevivência e longitudinais independentes, teve-se em conta a associação esperada entre a progressão do marcador tumoral com a sobrevivência dos pacientes e, como tal, procedeu-se a uma modelação conjunta destes dois processos para inferir sobre a associação destes, adotando a metodologia de efeitos aleatórios. Resultados indicam que a progressão longitudinal de CA15-3 está significativamente associada com a probabilidade de sobrevivência destes pacientes. Conclui-se que a análise independente devolve estimativas dos parâmetros enviesadas sendo necessário considerar a associação entre os dois processos num modelo conjunto de dados do cancro da mama.

Trabalho Financiado pela Fundação para a Ciência e Tecnologia (Bolsa de Doutoramento SFRH/BD/74166/2010)

Ana Borges, **galardoada com o Prémio SPE 2015**, licenciou-se em Matemática (ramo via ensino) pela Universidade Portucalense. É Mestre em Ensino da Matemática pela Faculdade de Ciências da Universidade do Porto sob a supervisão da Professora Doutora Maria de Fátima Carvalho e o Professor Doutor Paulo Santos. Encontra-se a terminar o doutoramento em Ciências, especialidade Matemática, na Universidade do Minho, sob a orientação da Professora Doutora Inês Sousa. É docente equiparado a assistente na Escola Superior de Tecnologia e Gestão de Felgueiras do Instituto Politécnico do Porto.

• **Prémio Carreira SPE -** *um testemunho*

No XXII Congresso, em Olhão, a SPE atribuiu o "Prémio Carreira – SPE", na sua 2ª edição, à **Professora M. Antónia Amaral Turkman**

em reconhecimento pelas suas relevantes contribuições no desenvolvimento científico, pedagógico e de divulgação da Estatística em Portugal.

Na breve e sucinta apresentação do seu percurso académico pretendeu-se, por um lado destacar a sua diversificada atividade como docente e investigadora e por outro lado, fazer uma homenagem de reconhecimento e gratidão pelo seu papel fulcral no desenvolvimento da Estatística e em particular da Estatística bayesiana em Portugal. Foi dado destaque à variedade de tópicos de investigação, cobrindo a Bioestatística, aplicações da Estatística no Ambiente, nas Ciências da Saúde e na Epidemiologia, assim como às diferentes metodologias usadas, de índole mais teórico ou metodológico com vista às aplicações. Em suma ficou salientada a sua sólida e abrangente formação estatística e a sua versatilidade e complementaridade no uso de ferramentas e metodologias estatísticas. Analogamente também foi referida a sua capacidade de realização multitarefa, como atestam a quantidade de publicações e de orientações pós-docs, de teses de doutoramento e de mestrado.

A Sociedade Portuguesa de Estatística também lhe está reconhecida pelo seu papel colaborante e efetivo na vida da SPE. Foi uma das onze personalidades outorgantes da criação da então designada Sociedade Portuguesa de Estatística e Investigação Operacional (28 de novembro de 1980) e por diversas vezes fez parte dos corpos sociais, sendo vice-presidente em dois mandatos.

Por fim, queria deixar registada a minha gratidão e reconhecimento à Prof Antónia. Foram determinantes na minha carreira a sua orientação científica e o seu apoio e total disponibilidade a cada momento do meu percurso.

Isabel Pereira
Vice-presidente da SPE

Edições SPE - Minicursos

Título: *Estatística Bayesiana Computacional – uma introdução*
Autores: M. Antónia Amaral Turkman e Carlos Daniel Paulino
Ano: 2015.

Título: *Análise de Valores Extremos: Uma Introdução*
Autoras: M. Ivette Gomes, M. Isabel Fraga Alves e Claudia Neves
Ano: 2013.

Título: *Modelos com Equações Estruturais*
Autora: Maria de Fátima Salgueiro
Ano: 2012.

Título: *Análise de Dados Longitudinais*
Autoras: Maria Salomé Cabral e Maria Helena Gonçalves
Ano: 2011

Título: *Uma Introdução à Estimação Não-Paramétrica da Densidade*
Autor: Carlos Tenreiro
Ano: 2010

Título: *Análise de Sobrevivência*
Autoras: Cristina Rocha e Ana Luísa Papoila
Ano: 2009

Título: *Análise de Dados Espaciais*
Autoras: M. Lucília de Carvalho e Isabel C. Natário
Ano: 2008

Título: *Introdução aos Métodos Estatísticos Robustos*
Autores: Ana M. Pires e João A. Branco
Ano: 2007

Título: *Outliers em Dados Estatísticos*
Autor: Fernando Rosado
Ano: 2006

Título: *Introdução às Equações Diferenciais Estocásticas e Aplicações*
Autor: Carlos Braumann
Ano: 2005

Título: *Uma Introdução à Análise de Clusters*
Autor: João A. Branco
Ano: 2004

Título: *Séries Temporais – Modelações lineares e não lineares*
Autoras: Esmeralda Gonçalves e Nazaré Mendes Lopes
Ano: 2003 (2ª Edição em 2008)

Título: *Modelos Heterocedásticos. Aplicações com o software Eviews*
Autor: Daniel Muller
Ano: 2002

Título: *Inferência sobre Localização e Escala*
Autores: Fátima Brilhante, Dinis Pestana, José Rocha e Sílvio Velosa
Ano: 2001

Título: *Modelos Lineares Generalizados – da teoria à prática*
Autores: M. Antónia Amaral Turkman e Giovani Silva
Ano: 2000

Título: *Controlo Estatístico de Qualidade*
Autoras: M. Ivette Gomes e M. Isabel Barão
Ano: 1999

Título: *Tópicos de Sondagens*
Autor: Paulo Gomes
Ano: 1998

Retrospetiva

O *Boletim SPE* através dos seus "Tema Central"

- Primavera de 2015 – Destaque: Estatística no Desporto
- Outono de 2014 – Destaque: Estatística no Ensino Básico e Secundário
- Primavera de 2014 – Destaque: (Um) Ano Internacional da Estatística
- Outono de 2013 – Destaque: A "Escola Bayesiana" em Portugal
- Primavera de 2013 – Destaque: Estatística não-paramétrica
- Outono de 2012 – Destaque: Métodos Estatísticos em Medicina
- Primavera de 2012 – Destaque: Estatística no Ensino Superior Politécnico
- Outono de 2011 – Destaque: Análise de Sobrevivência
- Primavera de 2011 – Destaque: Sondagens e Censos
- Outono de 2010 – Destaque: Estatística Espacial
- Primavera de 2010 – Destaque: Data Mining - Prospecção (Estatística) de Dados
- Outono de 2009 – Destaque: Modelos Econométricos
- Primavera de 2009 – Destaque: Investigação (em) Estatística
- Outono de 2008 – Destaque: Processos Estocásticos
- Primavera de 2008 – Destaque: ALEA - Um sítio do nosso mundo
- Outono de 2007 – Destaque: Bioestatística
- Primavera de 2007 – Destaque: A "Escola de Extremos" em Portugal
- Outono de 2006 – Destaque: Ensino e Aprendizagem da Estatística

também disponíveis em *http://www.spestatistica.pt/index.php/publicacoes-57/boletins*

SOCIEDADE PORTUGUESA
DE ESTATÍSTICA

www.spestatistica.pt

Índice