Modelagem matemática tempo-dependente de alterações genéticas na evolução da leucemia mieloide aguda

Matheus O Meirim[1], Juliana B da Costa[2],Luciana M Gutiyama[2], Ilana R Zalcberg[2], Julia L Fleck[1*]

[1]Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio). [2]Instituto Nacional do Câncer (INCA)

*Autor correspondente: jfleck@puc-rio.br

INTRODUÇÃO

A leucemia mieloide aguda (LMA) é uma neoplasia agressiva e a sobrevida em 5 anos é inferior a 30%. A heterogeneidade da LMA se reflete nessa alta mortalidade e dificulta avanços terapêuticos. Em geral, dados sobre lesões pré-leucêmicas se limitam a hierarquizar apenas alterações genômicas. Contudo, entende-se que mutações em genes *drivers* podem gerar mudanças de expressão de outros genes importantes para a progressão da LMA, e assim favorecer mutações adicionais em fases subsequentes da doença.

OBJETIVOS

Classificar hierárquica e temporalmente, em análise em corte transversal, mutações e mudanças de expressão gênica em pacientes com LMA ao diagnóstico via modelagem matemática.

MATERIAL E MÉTODOS

Foi utilizado um modelo de programação linear inteira mista (*Mixed Integer Linear Program* – MILP) para inferir a sequência temporal em que mutações somáticas e translocações ocorrem e geram mudanças na expressão gênica durante a progressão da LMA. O modelo MILP admite as hipóteses de que, ao longo de cada fase de evolução da doença, mutações somáticas em genes *drivers* geram mudanças de expressão em genes importantes para a progressão da LMA, e que alterações na expressão gênica podem favorecer mutações adicionais, bem como mudanças na expressão gênica de outros genes em fases subsequentes do desenvolvimento da doença. A formulação é baseada nos seguintes pressupostos:

1. Exclusividade de mutações: apenas um gene *driver* pode adquirir mutação somática em cada fase da progressão temporal do câncer.
2. Progressão de mutações através das fases: é necessário que exista uma mutação somática em algum gene em uma determinada fase para que outro gene possa sofrer uma mutação em uma fase subsequente.
3. Relação de dependência entre os mutações somáticas e mudanças de expressão gênica: a ocorrência de mutações somáticas em genes *drivers* desencadeia alterações na expressão gênica em genes importantes para a evolução do câncer. Tais mudanças de expressão, por sua vez, podem levar ao surgimento de novas mutações somáticas em fases posteriores da doença.
4. A força da conexão entre os genes com expressão alterada e os genes drivers com mutação somática determina a alocação dos genes com expressão anormal nas fases correspondentes. Neste contexto, entende-se por conexão o grau de relação entre a ocorrência de um evento mutacional em um determinado gene e a alteração na expressão de um outro gene.

Esse modelo é alimentado por duas matrizes. A primeira é uma matriz binária *M* onde cada linha (indexada por *i*) contém a identificação das amostras analisadas e cada coluna (indexada por *j*) é um gene que pode ter sofrido mutação ou translocação. O interior da matriz é preenchido linha a linha com o valor 0, caso o gene não tenha sofrido mutação na amostra e 1 caso tenha. Os valores M_{ij} da matriz de mutação são definidos da seguinte forma:

$$M_{ij} = \begin{cases} 1 & \text{if mutation gene } j \text{ is mutated in sample } i \\ 0 & \text{otherwise} \end{cases}$$

A segunda matriz *E*, com os dados de expressão, contém nas linhas (indexadas por *i*) os genes que podem ter sofrido alteração na expressão e nas colunas (indexadas por *h*) a identificação das amostras. O interior da matriz é preenchido coluna a coluna com os valores E_{ih} referentes à expressão do gene na amostra.

A partir dessas matrizes, é possível definir o conceito de conectividade entre os genes da matriz *M* e os genes da matriz *E*, que é mapeado através da matriz $C \equiv E^T \cdot M$. Os valores de conectividade gerados a partir desta operação matricial são alocados na matriz *C* que detém nas linhas (indexadas por *h*) os genes que sofreram alteração na expressão e nas colunas (indexadas por j) os genes que sofreram mutação ou translocação. O interior da matriz (C_{hj}) possui os valores de conectividade. Valores de conectividade próximos de 0 indicam que há uma conectividade fraca, enquanto que valores próximos de 1 indicam uma forte conectividade entre um determinado gene de expressão e um determinado gene de mutação ou translocação. Desta forma, percebe-se que uma forte conectividade decorre de um par de genes (*h,j*) que sofreu uma mutação e possui um valor de expressão muito acima da média dos valores encontrados na matriz *E*, respectivamente.

A partir desses dados de entrada, o modelo propõem a divisão dos genes das matrizes *M* e *E* em uma determinada quantidade de fases de progressão do câncer. Vale ressaltar que esse número de fases também é um parâmetro de entrada do modelo. A formulação do modelo inclui uma função objetivo (Equação (1)) e restrições (Equações (C1) – (C7)).

$$\min \left[\frac{1-W}{m \cdot n} \sum_{i=1}^{m} \sum_{k=1}^{K} \left(\sum_{j=1}^{n} M_{i,j} p_{j,k}^{M} - a_{i,k}^{M} + 2 f_{i,k}^{M} \right) - \frac{W}{K \cdot r} \sum_{k=1}^{K} \sum_{h=1}^{r} p_{h,k}^{E} \right] \quad (1)$$

$$\begin{aligned}
s.t.\ & \textstyle\sum_{k=1}^{K} p_{j,k}^{M} = 1 && \forall \text{ mutation gene } j && (C1)\\
& \textstyle\sum_{k=1}^{K} p_{h,k}^{E} \geq 0 && \forall \text{ expression gene } h && (C2)\\
& \textstyle\sum_{j=1}^{n} p_{j,k}^{M} \geq 1 && \forall \text{ phase } k && (C3)\\
& \textstyle\sum_{h=1}^{r} p_{h,k}^{E} \geq 0 && \forall \text{ phase } k && (C4)\\
& a_{i,k}^{M} \geq a_{i,k+1}^{M} && \forall \text{ sample } i, \forall \text{ phase } k && (C5)\\
& a_{i,k}^{M} \leq f_{i,k}^{M} + \textstyle\sum_{j=1}^{n} M_{i,j} \cdot p_{j,k}^{M} && \forall \text{ sample } i, \forall \text{ phase } k && (C6)\\
& p_{h,k}^{E} = \textstyle\sum_{j=1}^{n} C_{h,j} \cdot p_{j,k}^{M} && \forall \text{ expression gene } h, \forall \text{ phase } k && (C7)
\end{aligned}$$

O fluxograma da metodologia utilizada neste estudo está apresentado na Figura 1. O modelo foi alimentado com dados provenientes da base de dados LMA-TCGA. Os dados são disponibilizados em arquivos de amostras individuais contendo informações sobre cada paciente. Para o modelo utilizado, as informações relevantes são o número da amostra, bem como quais genes sofreram mutação, translocação e quais apresentaram alteração na expressão. A partir dos arquivos individuais, foram geradas duas matrizes de entrada do modelo (*M* e *E*), a primeira contendo os dados de mutação e translocação e a segunda os dados de expressão para cada amostra analisada.

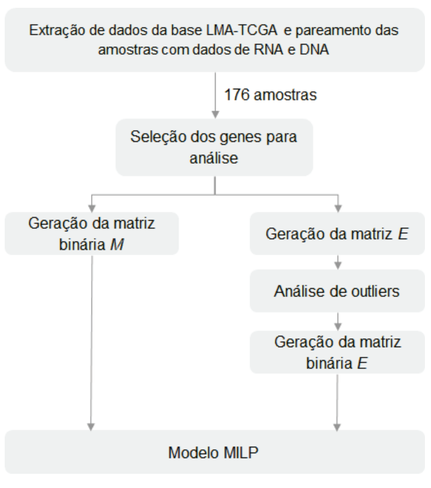

Figura 1 – Fluxograma da metodologia do estudo

Após as matrizes serem geradas, os dados passam por três fases de pré-processamento antes de serem alimentados ao modelo. Primeiramente, é feito um pareamento das amostras contidas nas duas matrizes, de forma que as duas matrizes contenham exatamente as mesmas amostras, retirando-se assim as amostras presentes apenas em uma das matrizes. A segunda fase do pré-processamento consiste em filtrar os genes contidos nas matrizes de mutação e expressão seguindo critérios de importância, obtidos da literatura, para o desenvolvimento de LMA. As duas primeiras fases de pré-processamento foram realizadas com auxílio computacional, a partir de scripts utilizando a linguagem Python. A terceira fase é exclusiva da matriz de expressão, e consiste em uma análise de outliers. A matriz de expressão E é pré-processada visando a determinação de quais genes serão mantidos de acordo com o percentil selecionado. O critério adotado na análise de outliers foi de considerar como sub-expressos os genes que estavam no percentil 1 e como sobre-expressos os genes que pertenciam ao percentil 99 da matriz E.

Por fim, o modelo foi alimentado com as duas matrizes de entrada, contendo um total de 176 amostras, 44 genes de mutação e 210 genes de expressão, e foram feitas rodadas variando-se o número de fases de progressão do câncer entre 2 e 5.

RESULTADOS

Estudos sobre a evolução clonal da LMA e, mais recentemente, dados sobre o envolvimento da CH com neoplasias hematológicas fornecem cada vez mais dados sobre a ordem de ocorrência de mutações ao longo da progressão da LMA, entretanto, dados sobre a expressão gênica ao longo das fases da LMA ainda são escassos. Assim, o modelo com 3 fases foi escolhido por demonstrar maior concordância com a literatura em relação aos genes mutados em cada fase. A Tabela 1 detalha os genes *drivers* mutados em cada uma das 3 fases de progressão da LMA propostas pelo modelo, enquanto a Tabela 2, os genes que tiveram suas expressões alteradas. A Figura 2 apresenta a quantidade de genes alocados a cada uma de 3 fases de evolução da LMA.

Tabela 1 – Genes *drivers* mutados a cada fase de progressão da LMA considerando-se uma divisão dos dados em 3 fases

Fase da LMA	Genes mutados	Total (N)
1	*DNMT3A, PTPN11, ASXL1, NRAS, U2AF1, WT1, JAK2, SF3B1, t(15;17), t(8;21)*	10
2	*NPM1, RUNX1, KDM6A, CEBPA, EZH2, TET2, MYC, SRSF2, U2AF2, ETV6, IKZF1, NF1, MLL2, BCOR, inv(3), inv(16), t(6;9), t(9;11), t(x;11), (x;q23)*	19
3	*IDH1, PHF6, IDH2, FLT3, TP53, RAD21, KIT, KRAS, STAG2, CBL, GATA2, SH2B3, MLL3, MPL, t(9;22)*	15

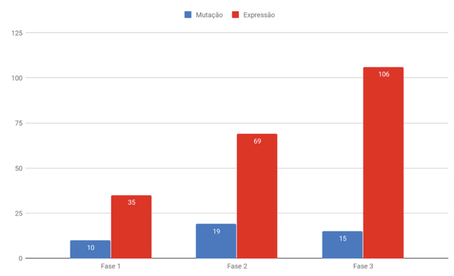

Figura 2 – Quantidade de genes alocados a cada fase de progressão da LMA considerando-se uma divisão dos dados em 3 fases

Tabela 2 - Genes com a expressão gênica alterada em cada fase de progressão da LMA considerando-se uma divisão dos dados em 3 fases

Fase da LMA	Genes com expressão alterada	Total (N)
1	*CDC42, ARNT, TPM3, AKT3, FAS, FGFR2, CCND1, CBL, BMP4, PLCB2, MAP2K1, PRKCB, TP53, GRB2, LAMA1, RALBP1, PRKACA, ROCK2, STAT1, RHOA, MITF, PLD1, FGFR3, MAPK10, NFKB1, VEGFC, SKP2, PDGFRB, RAC1, FGFR1, P,TK2, PRKACG, DAPK1, TGFBR1, WNT1*	35
2	*DVL1, PIK3CD, MTOR, CASP9, CSF3R, PRKACB, NRAS, FASLG, MAPK8, CTBP2, TRAF6, BAD, RELA, FGF19, FGF4, FADD, MMP1, ARHGEF12, CDKN1B, FLT3, BRCA2, FOXO1, NFKBIA, HIF1A, RAD51, AXIN1, CREBBP, MAPK3, PLCG2, STAT5B, STAT5A, BIRC5, ROCK1, SMAD4, BAX, KLK3, SOS1, RALB, GLI2, CASP8, GNAS, BID, CRKL, PDGFB, PPARG, MLH1, GSK3B, MECOM, KIT, PIK3R1, MSH3, APC, FGF1, MAPK9, PPARD, CDKN1A, HDAC2, CYCS, FZD1, SMO, BRAF, SHH, FGF20, MYC, GNAQ, RALGDS, ARAF, AR, CDH1*	69
3	*HDAC1, PTCH2, JUN, JAK1, NTRK1, TPR, PTGS2, TGFB2, FH, ITGB1, RET, CXCL12, NCOA4, CCDC6, PTEN, CHUK, NFKB2, SUFU, HRAS, BIRC3, BIRC2, ZBTB16, KRAS, CDK2, GLI1, CDK4, MDM2, HSP90B1, CCNA1, RB1, EDNRB, MAX, FOS, TGFB3, HSP90AA1, AKT1, DAPK2, SMAD3, PML, IGF1R, MMP2, NOS2, ERBB2, RARA, STAT3, AXIN2, PRKCA, SMAD2, BCL2, DAPK3, MAP2K2, PIK3R2, CCNE1, CEBPA, AKT2, TGFB1, PRKCG, MSH2, MSH6, CTNNA2, PAX8, CXCR4, ITGAV, FN1, STK36, BCL2L1, E2F1, PLCG1, MMP9, RUNX1, BCR, MAPK1, EP300, VHL, RAF1, RARB, TGFBR2, CTNNB1, TFG, PIK3CB, PIK3CA, CTBP1, PDGFRA, EGF, FGF2, EDNRA, CASP3, CSF1R, VEGFA, HSP90AB1, PDGFA, RALA, GLI3, EGFR, HGF, CDK6, MET, IKBKB, RUNX1T1, CDKN2A, PTCH1, ABL1, RXRA, TRAF2, XIAP, STK4*	106

Os genes listado na Tabela 1 foram classificados de acordo com sua função em oito categorias: metilação/ demetilação do DNA; componentes de vias de sinalização; modeladores de cromatina; fatores de splicing; fatores de transcrição mieloide; fatores de transcrição, componentes do complexo coesina e supressores de tumor e foram distribuídos ao longo da progressão da LMA (Figura 3).

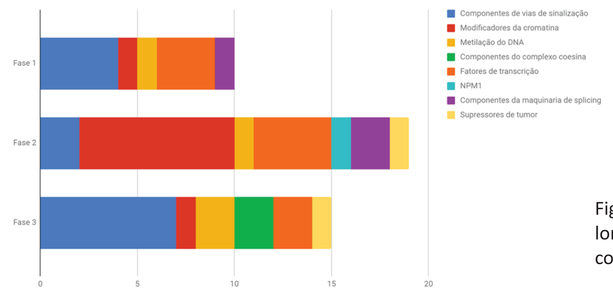

Figura 3 - Distribuição das mutações ao longo da progressão da LMA de acordo com a função gênica

DISCUSSÃO

Na Fase 1, o modelo selecionou 10 lesões *drivers* em genes conhecidos em estágios pré-leucêmicos e hematopoiese clonal de potencial indeterminado, como por exemplo o gene DNMT3A, U2AF1, sendo a maioria dos genes pertencentes à classe de moduladores epigenéticos, além das translocações t(15;17) e t(8;21), ambas caracterizam casos de LMA da classe core *binding factor* (CBL) e foram descritas como lesões pré-leucêmicas com necessidade de um segundo evento para a transformação maligna da LMA. Na Fase 2, foram incluídos 19 genes como NPM1, RUNX1 e CEBPA, estes utilizados para a classificação de risco da LMA. A maioria dos genes mutados nesta fase apresentam comprometimento com a diferenciação mieloide. Na Fase 3, foram selecionados 15 genes, dentre eles, FLT3 e TP53, em geral, detectados em fases tardias da LMA. A progressão da LMA se correlacionou proporcionalmente com o aumento da alteração na expressão gênica: 35 genes na Fase 1, 69 genes na Fase 2 e 106 genes na Fase 3. A alteração na Fase 1 se dividem entre genes codificadores de fatores de transcrição, citoesqueleto, GTPases e ciclinas, a maioria envolvida em vias de divisão e proliferação celular; a Fase 2 continua selecionando genes envolvidos com proliferação, entretanto somam-se genes envolvidos em vias apoptóticas e vias mais específicas da diferenciação mieloide e do câncer; a Fase 3 soma novas alterações às vias anteriores e inclui genes supressores de tumor.

CONCLUSÃO

O modelo, ainda que numa análise de corte transversal, foi capaz de hierarquizar mutações de acordo com a literatura e, de modo não supervisionado, a organização temporal das mudanças de expressão agrupou genes de mesma via. Assim, as ondas de expressão ocorreriam nesta ordem: 1) eventos epigenéticos e proliferação; 2) comprometimento da diferenciação mieloide e da apoptose; 3) expressão diferencial de genes supressores de tumor (Figura 4).

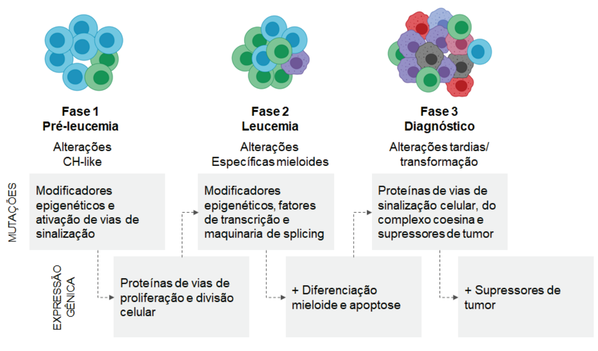

Figura 4 - Hipótese sobre a progressão da LMA baseada na aplicação do modelo matemático MILP

Projeto Gráfico: Área de Edição e Produção de Materiais Técnico-Científicos / INCA