Aprendizado de Máquina Aplicado ao Rastreamento de Partículas e Classificação de Modos de Difusão

Tiago Nascimento de Azevedo - 83085

29 de junho de 2022

Resumo

Neste artigo descrevemos e avaliamos algumas aplicações de técnicas de aprendizado de máquina em dados de difusão de partículas. Foram descritos seis artigos de quatro técnicas diferentes, como Random Forest (RF), Support Vector Machine (SVM), Redes Neurais Artificiais de Retropropagação (BPNN) e Redes Neurais Convolucionais (CNN).

1 Introdução

Processos difusivos são onipresentes na natureza e sua correta caracterização provê importante entendimento dos processos físicos associados à estes. A migração celular, por exemplo, tem um papel fundamental em muitos processos fisiológicos, como cicatrização de feridas, câncer e desenvolvimento de órgãos [HF13, WVS+13]. Assim, uma descrição adequada do movimento celular é essencial para a compreensão desses processos. Através da técnica da microrreologia, uma análise da caminhada aleatória de uma partícula de prova pode fornecer as propriedades mecânicas do material no qual a mesma difunde [Wai05]. Com uma única trajetória pode-se inferir propriedades físicas locais do meio, como o coeficiente de difusão e a densidade de obstáculos.

Várias técnicas foram inventadas no intuito de obter e/ou analisar trajetórias de partículas ou de um ensemble de partículas na escala microscópica/nanoscópica, como a espectroscopia de onda difusa (DWS), espalhamento dinâmico de luz (DLS) e a microscopia de rastreamento de partículas (PTM) [SM10]. De utilização comum à todos os trabalhos citados aqui, a técnica de rastreamento de partícula única (SPT) se destaca como uma poderosa e popular técnica para estudar processos dinâmicos nos mais diversos sistemas [MGP15].

A caracterização destas trajetórias incide basicamente na tarefa de classificação de seu modo de difusão que, tipicamente, se apresenta de diferentes quatro maneiras: difusão normal (movimento browniano livre), movimento direcionado (DM), difusão anômala (AD) e difusão confinada (CD). Usualmente isso é feito olhando-se para o deslocamento quadrático médio (MSD) das partículas, onde o expoente α em $\langle \Delta r^2(\tau) \rangle \propto t^\alpha$ fornece um indicativo do modo de difusão.

Em muitas situações as trajetórias são muito curtas para permitir uma análise através do MSD e técnicas de aprendizado de máquina podem ser utilizadas para superar essa limitação natural de alguns sistemas físicos. Não apenas para isso, algoritmos de aprendizado de máquina podem ser também utilizados para a segmentação destas trajetórias (através de técnicas de deep learning) [NSL+18]. Na seção seguinte apresentamos diversos trabalhos dedicados a aplicar todo o ferramental do aprendizado de máquina para a classificação de trajetórias de partículas nos mais variados sistemas. Por último, apresentamos as conclusões.

2 Aprendizado de Máquina Aplicado ao Rastreamento de Partícula Única e Classificação de Modos de Difusão

2.1 *Random Forest* (RF)

Árvores de decisão são estruturas de processamento de dados (semelhantes à árvores) utilizadas para problemas de classificação e regressão, onde cada nó representa uma condição ou regra e cada ramificação representa um possível resultado derivado desta condição (Fig. 1).

Random Forest (RF) é uma técnica de aprendizado supervisionado para regressão, classificação e outras tarefas, onde o resultado obtido através de várias árvores de decisão (criadas de maneira aleatória) são combinados para gerar um único resultado final (Fig. 1).

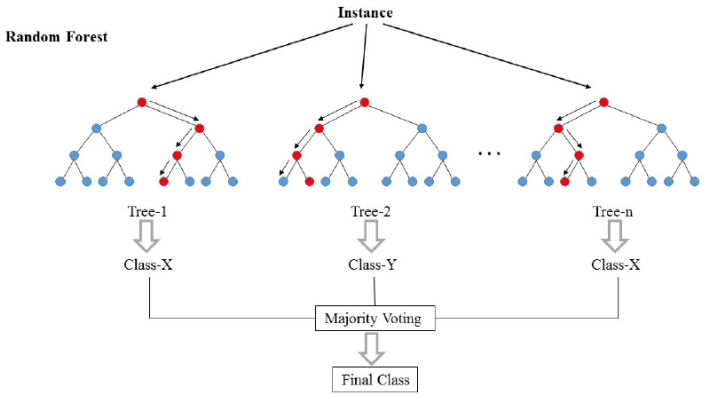

Figura 1: Ilustração do funcionamento da técnica de RF. Fonte [GSM+20].

Em [WKH+17] foi utilizada a técnica de RF para classificar trajetórias únicas de nanopartículas em microambientes celulares entre os seguintes tipos de difusão: normal, anômala, confinada e direcionada. Para a classificação foram consideradas as seguintes 9 classes normalizadas: O expoente anômalo alfa, assimetria, eficiência, dimensão fractal, gaussianidade, curtose, razão de deslocamento quadrático médio, retitude e aprisionamento. Aplicando o modelo em 20000 trajetórias (5000 de cada classe) bidimensionais (2D) obtidas via simulações de Monte Carlo foi obtida uma acurácia de 92%. Na Fig. 2 mostramos a matriz de confusão obtida. Com a mesma motivação, em [MGGMM+20] os pesquisadores

		Prediction			
		Directed	**Anomalous**	**Confined**	**Normal**
Reference	Directed	89.44%	0%	0.08%	10.48%
	Anomalous	0%	98.1%	1.34%	0.56%
	Confined	0%	2.88%	94.5%	2.62%
	Normal	6.36%	1.08%	3.14%	89.42%

Figura 2: Matriz de confusão do classificador RF aplicado em trajetórias sintéticas. Fonte [GSM+20].

utilizaram uma RF com 100 árvores para determinar o modo de difusão correspondente a uma dada trajetória. Com $1,2.10^5$ trajetórias simuladas (onde 80% destas é destinada ao treino e 20% é dedicada ao teste) o algoritmo foi capaz de classificar com ótima acurácia uma dada trajetória entre os modos CTRW, FBM e ATTM (*annealed transient time model*). A acurácia em função do comprimento da trajetória ($t_{\max}$) é mostrada na Fig. 3.

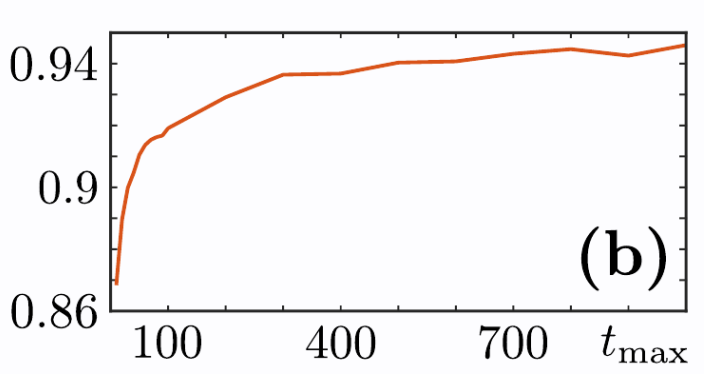

Figura 3: Acurácia do modelo em função do comprimento da trajetória $t_{\max}$. Fonte [MGGMM+20].

Como uma demonstração da funcionalidade do algoritmo, os pesquisadores o aplicaram a três

datasets, onde um é originado de dados simulacionais (modelos de compartimento, associados ao modo CTRW), outro (*Experiment* 1 em 4) corresponde a experimentos com moléculas de mRNA se movendo no interior de células bacterianas vivas (associado à FBM) e o terceiro (*Experiment* 2 em 4) vem da difusão de um receptor de membrana em células vivas (associado ao modelo ATTM). Na Fig. 4 é mostrada a porcentagem de trajetórias classificadas pelo algoritmo associadas a cada modelo.

Dataset	Predicted Model		
	CTRW	FBM	ATTM
(i) Compartments model	**89.2%**	0	10.7%
(ii) Experiment 1	4.5%	**86.6%**	8.9%
(iii) Experiment 2	16.4%	33.2%	**50.4%**

Figura 4: Porcentagem de trajetórias classificadas pelo algoritmo de RF associadas a cada modelo. Fonte [MGGMM+20].

2.2 *Support Vector Machine* (SVM)

Uma Máquina de Vetor de Suporte (SVM) é um algoritmo de aprendizado de máquina supervisionado utilizado para classificação e regressão de dados. Esta técnica visa encontrar o melhor hiperplano capaz de separar duas classes a partir dos dados de entrada (Fig. 5) [Vap95].

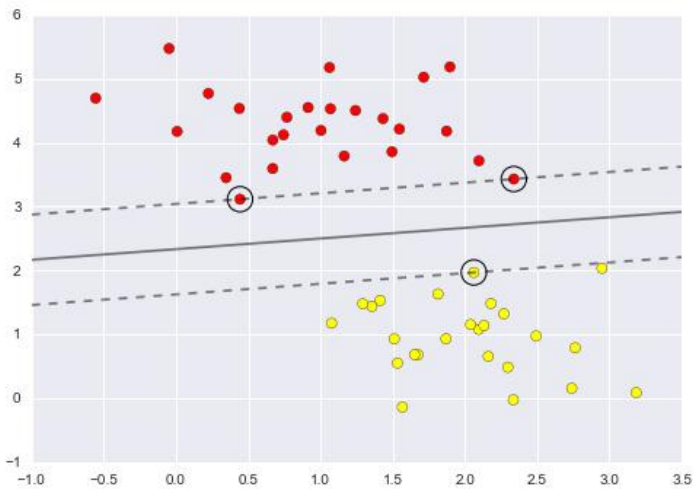

Figura 5: Esquema de classificador SVM ajustado aos dados (círculos coloridos) com margens (linhas tracejadas) e vetores de suporte (círculos coloridos destacados). Fonte [Van16].

Em [HBK+07] é desenvolvido um algoritmo baseado em SVM aplicado na identificação e classificação de trajetórias de partículas de Adenovirus-2 (Ad2) humano em células vivas. Baseada em 9 features (deslocamento líquido, 'straightness', 'bending', eficiência, assimetria, inclinação da posição do ponto e curtose da posição do ponto). a técnica foi treinada com dados sintéticos (simulações) e validada em experimentos, onde foi capaz de diferenciar quatro padrões de movimento: movimento confinado, deriva lenta, deriva rápida e movimento direcionado. As taxas de detecção para as diferentes trajetórias sintéticas são mostradas na Fig. 6 (Sens.: Sensitividade, Spec.: Especificidade).

	# steps	Sens.	Spec.
directed	2921	91.3 %	99.9 %
fast drift	8805	97.4 %	99.8 %
slow drift	133929	94.5 %	99.4 %
confined	311218	95.4 %	97.2 %
not classified	142944	95.1 %	96.4 %

Figura 6: Taxas de detecção do algoritmo baseado em SVM para diferentes tipos de movimento em trajetórias simuladas. Fonte [HBK+07].

2.3 *Back-Propagation Neural Network* (BPNN)

Inspiradas no cérebro humano, Redes Neurais Artificiais são uma técnica de aprendizado de máquina que consistem de camadas interconectadas de unidades de processamento (nós, ou neurônios), como ilustrado na Fig. 7. Sua estrutura consiste basicamente de uma camada de neurônios de entrada (que recebe os dados como input), uma camada de saída (que fornece o resultado final) e uma ou mais camadas intermediárias. No trabalho citado aqui, todos os nós de uma camada estão conectados a todos os nós das camadas adjacentes por interconexões chamadas sinapses (Fig. 7).

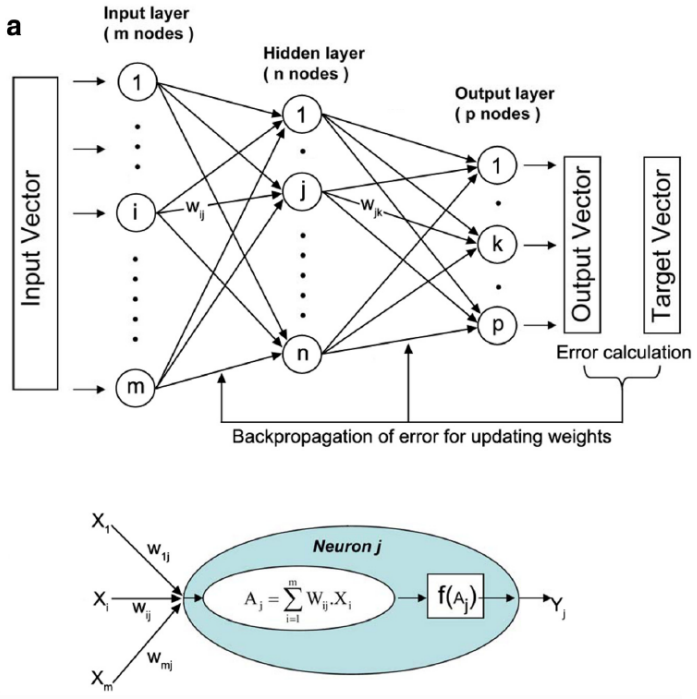

Figura 7: Acima: Arquitetura de uma rede neural de três camadas de retropropagação; Abaixo: representação de uma unidade de processamento (neurônio j), que recebe um vetor X_i como input e, com base na função de ativação $f(A_j)$ e nos pesos W_{ij} associados, fornece Y_j como output. Fonte [DRF+16].

Em uma rede neural do tipo *Back-Propagation*, com base no erro obtido na camada de saída, recalcula-se o valor dos pesos W_{ij} da última camada, retrocedendo e atualizando os pesos das camadas anteriores até a camada de entrada.

Em [DRF+16], uma rede neural de três camadas do tipo BPNN foi treinada e validada com 3000 trajetórias simuladas com 3.1 s de duração cada (1000 para cada modo de difusão: Browniano, confinado e direcionado). Como uma prova de conceito do método, para a validação, 200 trajetórias

Hiperparâmetro	Valor
Taxa de Regularização	0.0014064205292043147
Número de camadas convolucionais	6
Número de filtros	[49, 36, 18, 83, 90, 27]
Taxa de Aprendizado	0.00021795428728036654
Nós escondidos em camadas densas	1550

Tabela 1: Hiperparâmetros do modelo de rede com melhor desempenho mostrado.

simuladas de 400 quadros contendo um segmento de 50 quadros de movimento confinado e direcionado foram utilizadas. O algoritmo foi capaz de detectar os respectivos modos de difusão com uma boa acurácia (99% e 78% para os modos confinado e direcionado, respectivamente). O algoritmo foi também aplicado em trajetórias reais de células vivas, onde se mostrou capaz de detectar, com bastante acurácia, os regimes de difusão confinado e direcionado em segmentos de trajetórias.

2.4 *Convolutional Neural Network* (CNN)

Uma Rede Neural Convolucional é um algoritmo de Aprendizado Profundo (deep learning) utilizado principalmente em tarefas de classificação e processamento de imagens. Tomando uma imagem de entrada, o algoritmo atribui pesos a várias características da imagem (através da aplicação de filtros) e os diferencia. A Fig. 8 mostra um exemplo de uma arquitetura de uma CNN. Por operarem diretamente com os dados crus, métodos baseados em deep learning não querem qualquer preprocessamento complexo de dados ou extração de features por humanos.

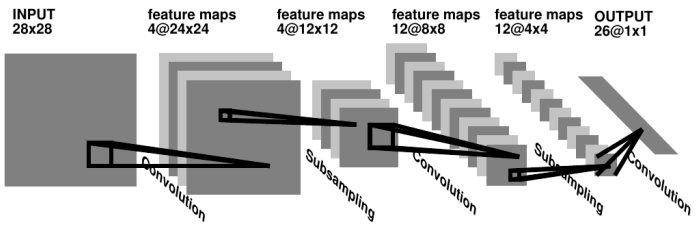

Figura 8: Exemplo de Rede Neural Convolucional para processamento de imagem. Fonte [BL97].

Em [KLOSan19] os autores abordam a tarefa de classificação de modelos de difusão de determinadas trajetórias utilizando uma CNN. Para treinar e validar o classificador, 20000 trajetórias sintéticas bidimensionais (correspondendo a diferentes tipos de modelos de difusão) geradas através de simulações de Monte Carlo foram utilizadas. Uma busca aleatória no espaço dos hiperparâmetros foi feita para determinar a melhor arquitetura da rede (resultado na Tabela 1).

A Fig. 9 mostra a matriz de confusão obtida, onde observamos uma excelente acurácia de 97.30%. Em [GWN+19] os autores desenvolvem um conjunto de redes neurais convolucionais para classificar trajetórias únicas dentre os 3 modos de difusão: movimento browniano, movimento browniano fracionário (FBM) e caminhada aleatória de tempo contínuo (CTRW). A rede neural foi treinada com 300000 trajetórias simuladas de 100 passos cada e validada experimentalmente em três situações (correspondendo cada uma a um modo de difusão): difusão de partículas fluorescentes em redes de actina (FBM), difusão de partículas fluorescentes em solução de água-glycerol (movimento browniano puro) e difusão de proteínas em superfícies de membrana (combinação de FBM e CTRW). Os resultados obtidos nos dados experimentais são mostrados na Fig. 11.

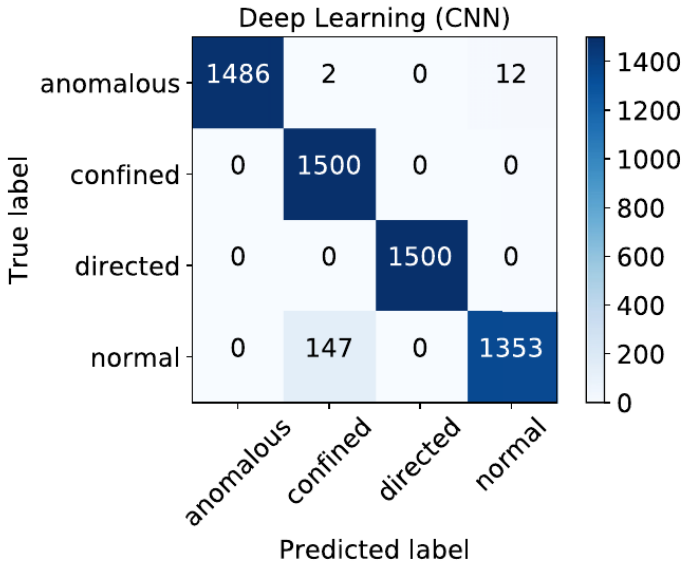

Figura 9: Matriz de confusão do classificador CNN. Fonte [KLOSan19].

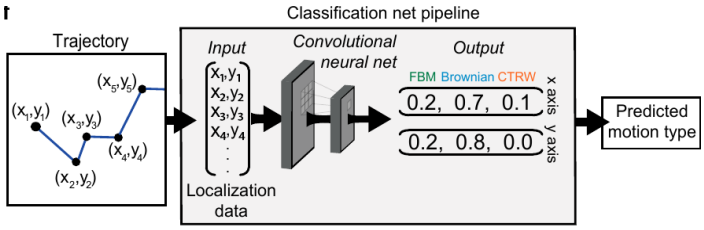

Figura 10: Uma representação esquemática do classificador. Fonte [GWN+19].

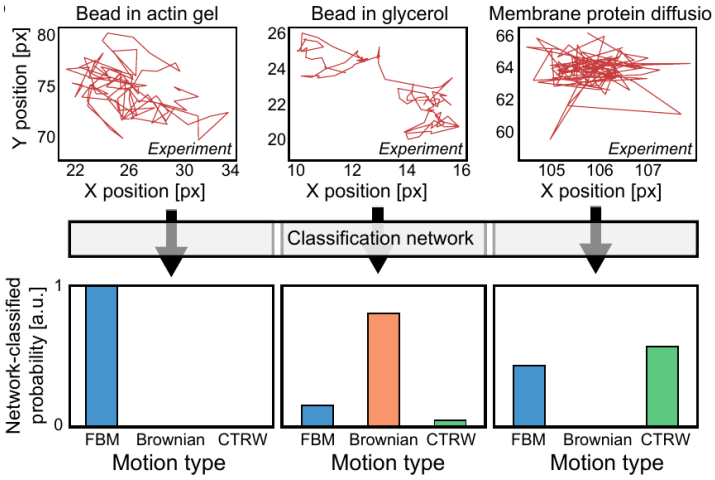

Figura 11: O resultado da análise da rede neural para três trajetórias medidas experimentalmente é mostrado; da esquerda para a direita: uma esfera fluorescente difundindo em um gel de actina demonstrando FBM; um grânulo fluorescente difundindo em uma mistura de glicerol-água, demonstrando movimento browniano; e difusão de proteína em uma membrana de célula viva exibindo uma combinação de FBM e CTRW. Fonte [GWN+19].

3 Conclusões

Neste trabalho foram apresentadas diversas técnicas de aprendizado de máquina aplicadas ao monitoramento e classificação de trajetórias únicas de partículas. Os algoritmos apresentados se revelaram bastante funcionais para essas tarefas, apresentando uma alta taxa de acerto. Ao custo de maior tempo de processamento dos dados, algoritmos de deep learning se mostraram ligeiramente superiores quando comparados ao resto.

Referências

[BL97] Y. Bengio and Yann Lecun. Convolutional networks for images, speech, and time-series. 11 1997.

[DRF+16] Patrice Dosset, Patrice Rassam, Laurent Fernandez, Cedric Espenel, Eric Rubinstein, Emmanuel Margeat, and Pierre-Emmanuel Milhiet. Automatic detection of diffusion modes within biological membranes using back-propagation neural network. *BMC Bioinformatics*, 17(1), 2016.

[GSM+20] Omid Ghorbanzadeh, Hejar Shahabi, Fahimeh Mirchooli, Khalil Valizadeh Kamran, Samsung Lim, Jagannath Aryal, Ben Jarihani, and Thomas Blaschke. Gully erosion susceptibility mapping (gesm) using machine learning methods optimized by the multi-collinearity analysis and k-fold cross-validation. *Geomatics, Natural Hazards and Risk*, 11(1):1653–1678, 2020.

[GWN+19] Naor Granik, Lucien E. Weiss, Elias Nehme, Maayan Levin, Michael Chein, Eran Perlson, Yael Roichman, and Yoav Shechtman. Single-particle diffusion characterization by deep learning. *Biophysical Journal*, 117(2):185–192, 2019.

[HBK+07] Jo A. Helmuth, Christoph J. Burckhardt, Petros Koumoutsakos, Urs F. Greber, and Ivo F. Sbalzarini. A novel supervised trajectory segmentation algorithm identifies distinct types of human adenovirus motion in host cells. *Journal of Structural Biology*, 159(3):347 – 358, 2007.

[HF13] Felix Höfling and Thomas Franosch. Anomalous transport in the crowded world of biological cells. *Reports on Progress in Physics*, 76(4):046602, mar 2013.

[KLOSan19] Patrycja Kowalek, Hanna Loch-Olszewska, and Janusz Szwabiński. Classification of diffusion modes in single-particle tracking data: Feature-based versus deep-learning approach. *Phys. Rev. E*, 100:032410, Sep 2019.

[MGGMM+20] Gorka Muñ oz-Gil, Miguel Angel Garcia-March, Carlo Manzo, José D Martín-Guerrero, and Maciej Lewenstein. Single trajectory characterization via machine learning. *New Journal of Physics*, 22(1):013010, jan 2020.

[MGP15] Carlo Manzo and Maria F Garcia-Parajo. A review of progress in single particle tracking: from methods to biophysical insights. *Reports on Progress in Physics*, 78(12):124601, oct 2015.

[NSL+18] Jay M. Newby, Alison M. Schaefer, Phoebe T. Lee, M. Gregory Forest, and Samuel K. Lai. Convolutional neural networks automate detection for tracking of submicron-scale particles in 2d and 3d. *Proceedings of the National Academy of Sciences*, 115(36):9026–9031, 2018.

[SM10] Todd M. Squires and Thomas G. Mason. Fluid mechanics of microrheology. *Annual Review of Fluid Mechanics*, 42(1):413–438, 2010.

[Van16] Jake VanderPlas. *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, Inc., 1st edition, 2016.

[Vap95] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

[Wai05] T A Waigh. Microrheology of complex fluids. *Reports on Progress in Physics*, 68(3):685–742, feb 2005.

[WKH+17] Thorsten Wagner, Alexandra Kroll, Chandrashekara R. Haramagatti, Hans-Gerd Lipinski, and Martin Wiemann. Classification and segmentation of nanoparticle diffusion trajectories in cellular micro environments. *PLOS ONE*, 12(1):1–20, 01 2017.

[WVS+13] Michael C. Weiger, Vidya Vedham, Christina H. Stuelten, Karen Shou, Mark Herrera, Misako Sato, Wolfgang Losert, and Carole A. Parent. Real-time motion analysis reveals cell directionality as an indicator of breast cancer progression. *PLOS ONE*, 8(3):1–12, 03 2013.