

# Caracterização dos Programas de Pós-graduação em Bioinformática no Brasil

Raquel C. de Melo-Minardi<sup>1</sup> Luciano A. Digiampietri<sup>2</sup>,  
Pedro O. S. Vaz de Melo<sup>1</sup>, Geraldo R. Franciscani Jr.<sup>1</sup>, Leonardo B. Oliveira<sup>1</sup>

<sup>1</sup>Departamento de Ciência da Computação da Universidade Federal de Minas Gerais

<sup>2</sup>Escola de Artes, Ciências e Humanidades da Universidade de São Paulo

{raquelcm,olmo,gfrancis,leob}@dcc.ufmg.br, digiampietri@usp.br

**Abstract.** *Multidisciplinarity is increasingly present in Science. The huge amount of available data as well their complexity have been demanding a large deal of effort from different areas for the understanding and resolution of problems. On the other hand, the interaction between people with different disciplinary education - alone - is a defying field. This paper describes the roles of the researchers in Brazilian post-graduate programs on Bioinformatics – an essentially multidisciplinary area – characterizing their interesting areas and collaboration networks.*

**Resumo.** *A multidisciplinaridade está cada vez mais presente na Ciência. A grande quantidade de dados disponíveis bem como a complexidade dos mesmos têm requerido esforços de diferentes áreas para o entendimento e a resolução de problemas. Por outro lado, a interação entre pessoas com diferentes formações disciplinares por si já traz novos desafios. Este artigo descreve a atuação dos pesquisadores nos programas brasileiros de pós-graduação em Bioinformática, uma área essencialmente multidisciplinar, caracterizando suas áreas de atuação e suas redes de colaboração.*

## 1. Introdução

A Bioinformática é uma área essencialmente interdisciplinar envolvendo as Ciências Biológicas (CB), Ciência da Computação (CC), Estatística, Química, Farmácia, Matemática, entre outras, no desenvolvimento de métodos para armazenamento e recuperação de dados biológicos e na construção de modelos e algoritmos para a solução de problemas biológicos. Embora se pense que a Bioinformática se originou com os projetos de sequenciamento de genomas a partir da década de 90, o termo foi cunhado no início da década de 70 por Ben Hesper visando a dar nome à pesquisa que ele desenvolvia e significando o estudo de processos informacionais em sistemas bióticos [Hogeweg 2011]. Contudo, algoritmos clássicos em Bioinformática já haviam sido desenvolvidos na década de 60 como, por exemplo, algoritmos para alinhamento de sequências e filogenia.

Segundo [Hogeweg 2011], o holandês Jaap Heringa foi o primeiro graduado em Bioinformática em 1984. No Brasil, os primeiros cursos a formar bioinformatas foram no nível de pós-graduação, especialmente de doutorado, e foram fomentados pelo edital Biomicro da CAPES em 2003. Esse edital foi uma solicitação dos participantes de um Seminário Nacional de Pós-Graduação e ofereceu 5 anos de suporte financeiro, incluindo bolsas e recursos de custeio. O objetivo era a criação de cursos de doutorado em

Bioinformática e em Microeletrônica. Nessa ocasião, a Universidade Federal de Minas Gerais (UFMG) e a Universidade de São Paulo (USP) apresentaram propostas e foram contempladas.

Esses programas, no momento de sua criação, reuniam pesquisadores das mais diversas áreas do conhecimento. Entretanto, os programas ficaram alocados na área de CB I. Posteriormente, foram criados o Programa de Pós-Graduação *stricto sensu* em Computação Biológica e Sistemas do Instituto Oswaldo Cruz em 2007 e o Programa de Pós-Graduação em Bioinformática da Universidade Federal de Paraná (UFPR) em 2009. Ambos fazem parte da área Interdisciplinar da CAPES.

A avaliação desses programas de pós-graduação é uma tarefa complexa devido principalmente à enorme diversidade de áreas de origem dos pesquisadores componentes do corpo docente. Como comparar pesquisadores que têm sua área de formação e de pesquisa em CB com os de CC quando existem tantas disparidades? Um exemplo é a questão da publicação de resultados em conferências ser valorizada em CC e nem tanto na maioria das outras áreas. Em Biologia, as publicações importantes são sempre feitas em periódicos indexados. Nesse cenário de tanta diversidade, como avaliar os programas de pós-graduação e os pesquisadores em termos de produtividade? Mais ainda, como avaliar os programas com diferentes composições de corpo docente em termos de áreas de origem?

Como os primeiros programas de pós-graduação em Bioinformática surgiram há 10 anos no Brasil, pouco se conhece sobre como os pesquisadores colaboram nesses programas criados para serem interdisciplinares. Segundo [Balancieri et al. 2005], a estrutura e a dinâmica da rede colaborações científicas, na forma de coautoria em produções bibliográficas, vem interessando às entidades avaliadoras e de fomento em ciência e tecnologia. No caso de Bioinformática, em se considerando o modelo usado pelo Lattes, essa rede é naturalmente segmentada em diferentes grandes áreas e áreas. Como essas áreas se comunicam na interdisciplinaridade da Bioinformática? Terá a criação dos programas de pós-graduação em Bioinformática realmente fomentado a criação de grupos interdisciplinares? Existe um padrão inerente a essas redes de colaboração? Como esses programas têm colaborado entre si?

**Objetivos.** O objetivo deste trabalho é caracterizar o conjunto de pesquisadores atuantes nos programas de pós-graduação em Bioinformática no Brasil. Visamos compreender importantes questões sobre como esses programas são compostos em termos de áreas de pesquisa dos docentes, como os pesquisadores dessas diferentes áreas podem ser comparados e contrastados considerando métricas de produtividade, como esses pesquisadores colaboram dentro e fora dos seus programas e como eles atuam de forma intradisciplinar ou interdisciplinar.

**Contribuição.** Mais precisamente, nossas principais contribuições são:

1. Caracterização da distribuição de grandes áreas e áreas dos pesquisadores atuantes nos programas de pós-graduação em Bioinformática;
2. Comparação entre médias de métricas tradicionais de produtividade docente tais como publicações de artigos, orientações concluídas e participação em bancas em diferentes áreas participantes dos programas;
3. Caracterização da evolução da rede de coautoria entre os pesquisadores dos pro-

gramas visando a compreensão de padrões e tendências de colaborações intra e inter programas bem como intradisciplinares e interdisciplinares.

**Organização.** O restante deste artigo está organizado da seguinte maneira. A Seção 2 discute os principais trabalhos correlatos. A Seção 3 apresenta a metodologia utilizada. A Seção 4 contém os resultados produzidos. Por fim, a Seção 5 descreve as conclusões e os trabalhos futuros.

## 2. Trabalhos Correlatos

Recentemente o número de trabalhos voltados para a análise de produtividade e interação entre pesquisadores, universidades e institutos de pesquisa cresceu substancialmente ([Bollen et al. 2006, Hirsch 2005, Garfield 2006, Mena-Chalco and Cesar-Jr. 2009, Martins et al. 2010, Duffy et al. 2011], por exemplo). A maioria deles utiliza métricas como o H-index [Hirsch 2005], o Fator de Impacto [Garfield 1955, Garfield 2006] e o número de citações em suas análises. Por exemplo, Martins et al. [Martins et al. 2010] avaliaram a qualidade de conferências baseando-se no número de citações. Eles chamaram a atenção para a necessidade de novas métricas de avaliação e propuseram então algumas particularmente voltadas a essa sorte de veículo de publicação.

Outros ([Menezes et al. 2009, Horn et al. 2004, Hayat and Lyons 2010, Franceschet 2011], por exemplo), por outro lado, empregam técnicas de análise de redes sociais (do inglês, SNA), comumente sobre a rede formada pela coautoria de trabalhos publicados. Por exemplo, Horn *et al.* [Horn et al. 2004] analisaram a evolução das redes de coautoria sobre o assunto “trabalho cooperativo apoiado pelo computador”, identificando áreas correlatas e inferindo padrões de colaboração.

Similarmente ao nosso trabalho, muitos restringem o objeto de estudo a uma área específica em certo país ou região. ([Cainelli et al. 2006, Rodríguez 2006, Guimaraes 2001, Laender et al. 2008, Freire and Figueiredo 2011, Digiampietri et al. 2012, Mena-Chalco et al. 2012], por exemplo). Dentre eles, Cainelli *et al.* [Cainelli et al. 2006], Rodríguez [Rodríguez 2006] e Guimarães [Guimaraes 2001] estudaram, respectivamente, a produção científica em economia da Itália, Espanha e Portugal. Acerca da área de CC no Brasil, em particular, Laender *et al.* [Laender et al. 2008] avaliaram a excelência dos principais programas de pós-graduação em CC no país. Eles contrastaram tais programas com programas renomados da América do Norte e Europa e concluíram que os programas brasileiros chegaram à maturidade. Figueiredo e Freire [Freire and Figueiredo 2011], por sua vez, estudaram a rede social acadêmica brasileira em CC. Em seu trabalho, eles observaram a existência de super pares (*peers*), ou seja, que um número restrito de pesquisadores possui uma grande quantidade de colaboradores quando comparado à grande maioria dos demais pesquisadores da rede. Por fim, Digiampietri *et al.* [Digiampietri et al. 2012] estudaram a dinâmica de relações de coautoria entre pesquisadores associados aos programas de pós-graduação em CC avaliados pela CAPES no triênio 2007-2009. Eles destacaram que através das redes de coautoria é possível observar interessantes fenômenos da dinâmica da pesquisa brasileira, em especial, aqueles relacionados ao aumento da produção conjunta inter e intraprogramas.

Acerca da área de Bioinformática, diversos trabalhos foram realizados no sentido de se caracterizar a pesquisa mundial

([Janssens et al. 2007, Bansard et al. 2007, Guan and Gao 2008, Glänzel et al. 2009, Jeong et al. 2009, Molatudi et al. 2009, Song and Kim 2012], por exemplo). Janssens *et al.* [Janssens et al. 2007], por exemplo, revelaram a estrutura da área de Bioinformática ao analisarem 7.401 publicações da *Web of Science*<sup>1</sup> e MEDLINE<sup>2</sup>. Para tal, eles empregaram um método de agrupamento híbrido, o que possibilitou a integração de dados textuais e citações e, por sua vez, uma melhora significativa nos resultados. Já Bansard *et al.* [Bansard et al. 2007], também apresentaram estudos com base no MEDLINE, mas agora à procura de tendências em Bioinformática. Eles apresentaram as palavras-chave mais recorrentes nos trabalhos do período 2000-2005; e contrastaram com a área de Informática Médica. Segundo eles, as áreas de Bioinformática e Informática Médica possuem linhas de pesquisa distintas, cabendo a elas compartilharem apenas os mesmos recursos, isto é, o mesmo aparato computacional e matemático. Guan *et al.* [Guan and Gao 2008] e Molatudi *et al.* [Molatudi et al. 2009], por sua vez, utilizaram dados do *Science Citation Index Expanded*<sup>3</sup> para caracterizar a área de Bioinformática na China e África do Sul, respectivamente. O primeiro grupo contrastou os resultados chineses com os dos EUA, Reino Unido, Japão e Índia. Ao passo que o segundo, com a Austrália, Brasil e Índia. Molatudi *et al.* também destacaram que, até 1998, Brasil e África do Sul caminhavam juntos na fatia de produção científica mundial na área (ambos com cerca de 0,5% da publicação mundial), mas, subsequentemente, até 2006, ocorreu um descolamento da produção dessas nações, com o Brasil aumentando sua fatia para cerca de 1,2% e a África apenas mantendo a sua fatia de 0,5%. Por fim, mais recentemente, Song *et al.* [Song and Kim 2012] publicaram outro estudo acerca da Bioinformática. Ao contrário da maioria dos demais trabalhos, os autores não basearam sua análise em dados bibliométricos. Eles concentram-se nos dados obtidos através da mineração de texto de artigos da *PubMed Central*<sup>4</sup> publicados na área. Seus resultados mostraram que há um crescimento linear na produção científica da área. A pesquisa também revelou que, dentro da Bioinformática, os tópicos mais recorrentes abordam mais os aspectos biológicos que os computacionais.

Cumpra lembrar que diferentemente do trabalho em questão, nenhum dos trabalhos supracitados delineiam a área da Bioinformática no Brasil.

### 3. Metodologia

A metodologia utilizada neste artigo foi dividida em: identificação dos dados de interesse, obtenção e organização dos dados, e análise dos dados.

A identificação dos dados de interesse foi realizada manualmente através da busca do identificador do currículo Lattes de cada um dos 104 pesquisadores dos programas de Pós-Graduação em Bioinformática. O nome destes pesquisadores foi obtido do caderno de avaliação<sup>5</sup> da CAPES de cada um dos quatro programas de interesse, referente a avaliação realizada no triênio 2007-2009.

---

<sup>1</sup>[thomsonreuters.com/products\\_services/science/science\\_products/a-z/web\\_of\\_science/](http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/)

<sup>2</sup><http://en.wikipedia.org/wiki/MEDLINE>

<sup>3</sup>[http://thomsonreuters.com/products\\_services/science/science\\_products/a-z/science\\_citation\\_index\\_expanded/](http://thomsonreuters.com/products_services/science/science_products/a-z/science_citation_index_expanded/)

<sup>4</sup><http://www.ncbi.nlm.nih.gov/pmc/>

<sup>5</sup><http://conteudoweb.capes.gov.br/conteudoweb/>

Com a lista de identificadores, cada currículo foi baixado, um banco de dados foi criado e as coautorias foram identificadas utilizando-se a metodologia apresentada em [Digiampietri et al. 2012]. Adicionalmente, todos os currículos Lattes referenciados através de endereços html em cada um dos 104 currículos também foram baixados e armazenados em um banco de dados utilizando a mesma metodologia. Este processo identificou 4.091 currículos que chamamos currículos dos colaboradores.

Para a análise de dados foram calculadas médias, correlações entre os atributos de produtividade de cada um dos pesquisadores bem como dos pesquisadores agrupados por área, grande área de atuação e por programa de pós-graduação. Adicionalmente, foram construídas redes de coautorias seguindo esse mesmo agrupamento. Técnicas de análise de redes sociais foram utilizadas para comparar as redes formadas.

## **4. Resultados**

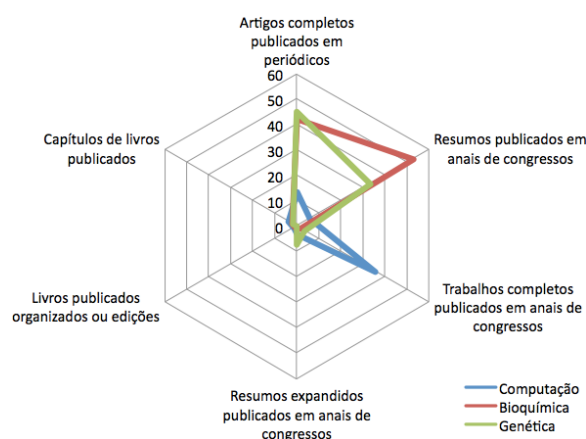
### **4.1. Composição dos programas**

Primeiramente, analisamos a composição dos programas de pós-graduação em Bioinformática em termos das grandes áreas e áreas dos 104 pesquisadores do corpo docente. Ressaltamos que cada pesquisador pode atribuir diversas grandes áreas e áreas nesta seção e, portanto, os valores deste gráfico somam um valor maior que 104. Como esperado, há mais menções à área de Ciências Biológicas (194) que é seguida pela área de Ciências Exatas e da Terra (123) e que somadas respondem por 86% das menções a grandes áreas. Quando analisamos as áreas, há três mais proeminentes: CC (88), Bioquímica (84) e Genética (46), respondendo por 59% das menções a áreas. As 10 áreas mais mencionadas (CC, Bioquímica, Genética, Biofísica (17), Probabilidade e Estatística (14), Engenharia de Produção (10), Imunologia (10), Química (10), Engenharia Biomédica (8), Microbiologia (8)) correspondem a 80% das menções. Ao todo, 32 diferentes áreas foram citadas. Quando analisamos a rede dos pesquisadores dos programas de Bioinformática juntamente com os seus vizinhos na rede de colaboradores externos aos programas, percebemos que há muito mais ligações com pesquisadores da grande área de Ciências Biológicas (53%).

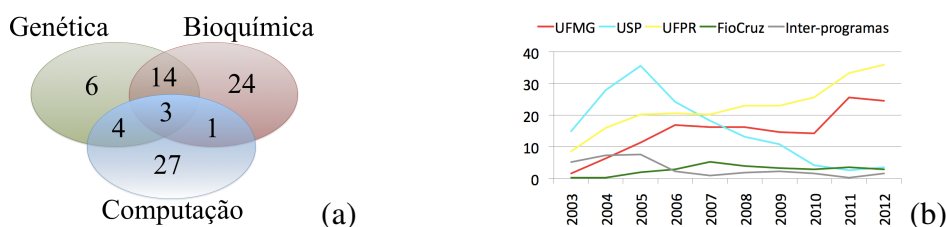
### **4.2. Métricas de produtividade das áreas mais representadas**

Como os programas contam com a participação de docentes de um número muito grande de áreas (32), a maioria com baixa representatividade, decidimos prosseguir analisando as três mais mencionadas: CC, Bioquímica e Genética. Em todos os programas, o cerne do corpo docente é composto majoritariamente pelas mesmas.

Sabe-se que os pesquisadores dessas áreas possuem estratégias de trabalho bastante diferentes. Essa diversidade inerente aos programas interdisciplinares faz com que sua comparação e avaliação sejam tarefas complexas. A Figura 1 apresenta múltiplas variáveis descritivas da produtividade média dos pesquisadores que se classificam nessas 3 áreas em termos de produção bibliográfica. Conforme esperado, a CC é a mais diferenciada tendo uma grande produção em anais de congressos e muito menor em periódicos. Nota-se ainda que a Bioquímica e a Genética valorizam mais os resumos ou pôsteres em conferências, o que não se verifica em CC. Embora esses sejam fatos conhecidos, eles são aqui quantificados pela primeira vez no cenário dos programas de pós-graduação brasileiros.



**Figura 1. Comparativo entre a produção científica média das áreas de CC, Bioquímica e Genética nos programas de Bioinformática.**

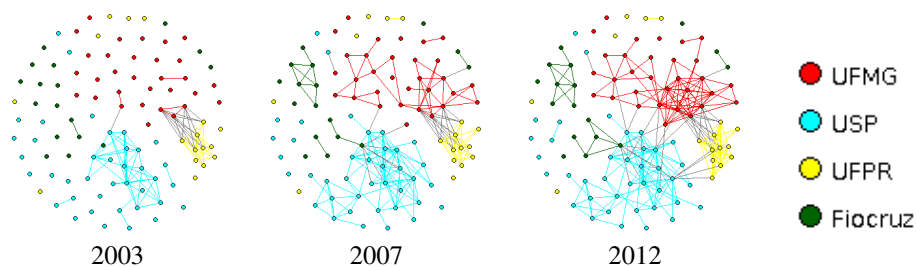


**Figura 2. (a) Distribuição dos 104 pesquisadores dos programas de pós-graduação em Bioinformática entre as três áreas mais representadas. (b) Média dos últimos três anos do número de publicações internas a cada um dos quatro programas e de publicações inter-programas.**

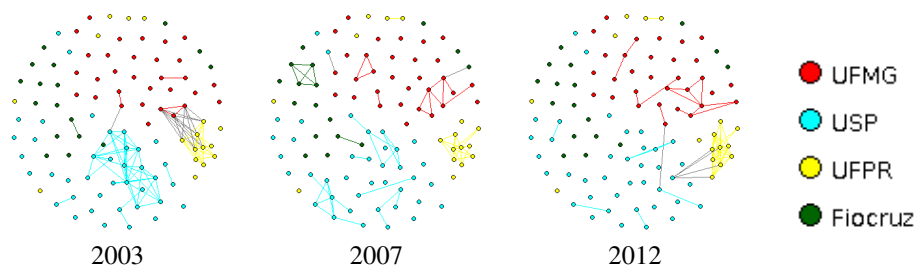
### 4.3. Redes de coautoria entre programas e entre áreas

Analizamos a seguir, como os pesquisadores dos diferentes programas têm colaborado desde 2003. É importante lembrar que os programas da UFMG e USP foram criados em 2003, o da FioCruz em 2007 e o da UFPR em 2009. A Figura 3 mostra a evolução acumulada das colaborações intra e inter-programas. É possível observar que os pesquisadores da USP e da UFPR já estabeleciam colaborações internamente desde 2003. Existia naquela época uma significativa colaboração entre membros do programa da UFMG com os da UFPR. Com o passar dos anos, o número de professores que já haviam colaborado cresceu intra-programa assim como entre os programas da UFMG e USP. A partir de 2011, nota-se que há grupos de professores fortemente conectados dentro dos programas da UFMG, USP e ainda que o grupo da UFPR era conectado como um todo. Em resumo, em 2012, uma parcela considerável dos docentes dos programas já havia estabelecido alguma colaboração intra ou inter-programas de forma que pode-se afirmar que há uma rede de conhecimento entre pesquisadores bastante conectada.

Contudo, quando analisamos a evolução e a manutenção dessas colaborações, o panorama não parece tão animador. Na Figura 4, notamos que as colaborações que existiam no programa da USP em 2003 tiveram uma grande redução. Na UFMG, a rede de colaborações tem uma conectividade muito variável ao longo dos anos, não havendo tendência evidente. A UFPR, entretanto, possui o único programa que permaneceu extremamente colaborativo nesses 10 anos de análise. A FioCruz ainda é um programa que possui poucas colaborações internas.



**Figura 3. Evolução da rede de coautoria acumulada nos programas.**



**Figura 4. Evolução da rede de coautoria nos programas.**

A Figura 2 (b) apresenta a média trienal do número de arestas dessas redes de coautoria no período de 2003 a 2012. Podemos perceber que o número de publicações inter-programas não passa de um quarto do valor total de publicações intra-programas, o que é natural dada a barreira geográfica existente entre os diferentes grupos de pesquisa. É possível quantificar a queda das colaborações internas no programa da USP e um crescimento bastante expressivo na UFPR. É notável como a UFPR apresenta uma curva de perfil similar à da UFMG mas sempre com maior magnitude mesmo tendo um corpo docente de tamanho (17) próximo à metade de UFMG (31) e USP (39) e comparável ao da Fiocruz (19). Isso indica um grupo de professores altamente coeso e colaborativo internamente. As colaborações inter-programas são extremamente variáveis no período e não pudemos verificar uma tendência específica.

A Figura 5 visa comparar as colaborações intradisciplinares com as interdisciplinares. Os nós representam os docentes e são posicionados na circunferência. As arestas são as coautorias e são dispostas cruzando o círculo. Agrupamos os docentes por programa, a seguir pela primeira grande área de atuação e, posteriormente, pela primeira área. As cores indicam as grandes áreas. A rede possui todas as colaborações estabelecidas entre 2003 e 2012. Nota-se claramente uma tendência de colaboração disciplinar (apresentadas em cores) em detrimento das interdisciplinares (em cinza). Mesmas as colaborações inter-programas são em sua maioria disciplinares. Uma exceção interessante é a USP, que apresentou um grande número de colaborações entre pesquisadores de Ciências Exatas e CB. A UFPR apresentou também colaborações significativas entre pesquisadores das Engenharias e CB. Por falta de espaço, não exibimos aqui dados quantitativos dessas colaborações embora eles possam ser visualizados em matrizes de adjacência interativas no endereço <http://each.uspnet.usp.br/digiampietri/brasnam2013/>.

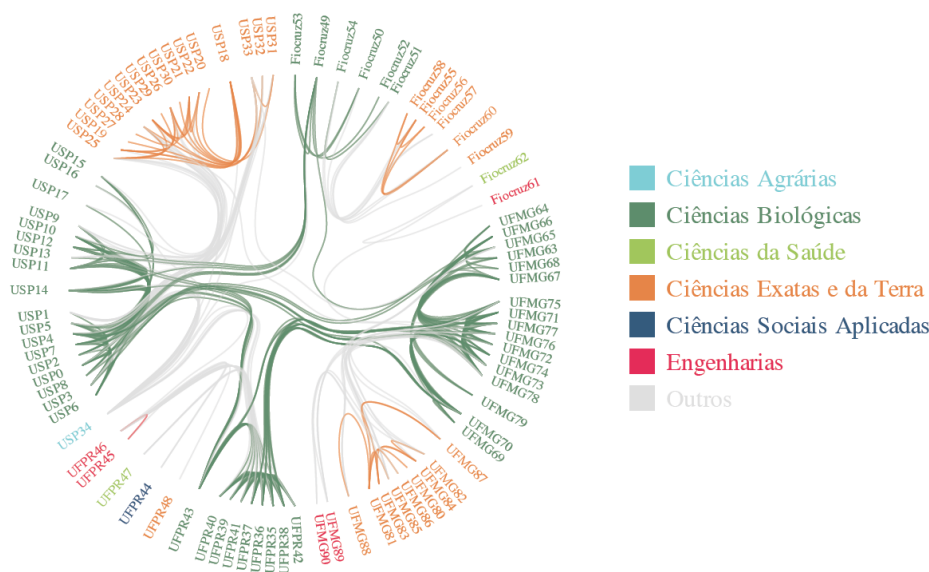


Figura 5. Rede de coautoria acumulada por grande área.

#### 4.4. Relações Intra e Interdisciplinares

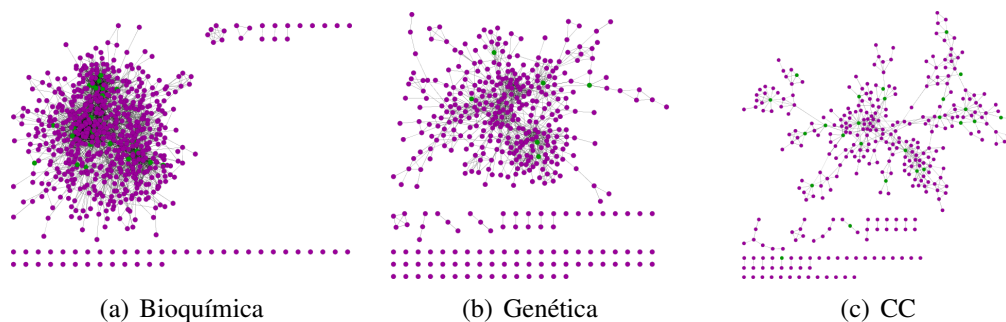
Uma das questões mais interessantes que se pode fazer quando um programa de pós-graduação multidisciplinar é analisado é: como são feitas as colaborações dentro e entre cada grupo de pesquisadores de uma determinada disciplina? No caso da Bioinformática, será que pesquisadores de CC se relacionam da mesma maneira que os de Bioquímica? As disciplinas têm participação similar nas produções acadêmicas ou há uma disciplina que se destaca quantitativamente? O objetivo desta seção é desconstruir a rede de colaborações dos programas de pós-graduação em Bioinformática a fim de entender melhor como são dadas as colaborações intra e interdisciplinares dentro da produção acadêmica desses programas.

Para fazer isso de forma detalhada, consideramos somente as três principais áreas declaradas do Lattes dos pesquisadores: Bioquímica, genética e CC, o que abrange 56% dos pesquisadores envolvidos em publicações ligadas à um programa de bioinformática. Assim, estamos considerando todos os pesquisadores que declararam uma dessas três áreas como a sua área de pesquisa e também todos aqueles que declaram mais de uma delas, por exemplo, “bioquímica e genética”.

Limitado o universo de pesquisadores e áreas de pesquisa, construímos grafos de colaboração compostos de pesquisadores que atuam somente em uma ou mais áreas. Na Figura 6 mostramos o grafo de colaboração de pesquisadores que atuam somente na área de Bioquímica (Figura 6 (a)), Genética (Figura 6 (b)) e CC (Figura 6(c)). Os nós verdes representam pesquisadores lotados em programas de pós-graduação em Bioinformática e os nós roxos, pesquisadores que colaboraram com os primeiros em pelo menos um dos dez anos em análise.

Primeiramente, note como os grafos de colaborações dos pesquisadores são significativamente diferentes entre si, principalmente quando comparamos o grafo daqueles que atuam em Bioquímica com o dos que atuam em CC. Enquanto o primeiro parece não ter estrutura bem definida, sendo similar a grafos aleatórios, o segundo parece ser mais





**Figura 6. Grafos de colaboração contendo pesquisadores que atuam em uma única área.**

**Tabela 1. Métricas globais dos grafos formados a partir da combinação das disciplinas bioquímica, genética e ciência da computação.**

| Grafo                       | $ V $ | $ E $  | $\bar{k}$ | $A_k$ | $A_{area}$ | $cc$ | $max(d_{i,j})$ | $avg(d_{i,j})$ |
|-----------------------------|-------|--------|-----------|-------|------------|------|----------------|----------------|
| Bioquímica                  | 719   | 3.156  | 8,78      | 0,00  | 1,00       | 0,48 | 8              | 3,55           |
| Genética                    | 435   | 1.245  | 5,72      | 0,26  | 1,00       | 0,42 | 13             | 4,29           |
| CC                          | 341   | 567    | 3,33      | 0,01  | 1,00       | 0,42 | 18             | 7,40           |
| Bioq. $\cup$ Gen.           | 1.516 | 13.266 | 17,50     | 0,23  | 0,17       | 0,55 | 10             | 3,34           |
| Bioq. $\cup$ CC             | 1.081 | 4.300  | 7,96      | 0,08  | 0,60       | 0,48 | 13             | 4,35           |
| Gen. $\cup$ CC              | 805   | 2.445  | 6,07      | 0,33  | 0,50       | 0,45 | 14             | 5,12           |
| Bioq. $\cup$ Gen. $\cup$ CC | 1.915 | 16.470 | 17,20     | 0,27  | 0,18       | 0,54 | 11             | 3,66           |

estruturado, contendo pequenas e distantes aglomerações de pesquisadores. Note também como os nós verdes estão mais distantes no grafo de CC, o que indica mais uma vez o baixo grau de colaboração entre pesquisadores de CC nos programas de Bioinformática. Por fim, pode-se notar que o grafo dos pesquisadores que atuam em Genética parece estar no meio do caminho entre o de bioquímica e o de CC.

A fim de entender melhor as observações supracitadas, listamos na Tabela 1 algumas das principais métricas globais de redes complexas para os grafos formados a partir das combinações das disciplinas declaradas pelos pesquisadores analisados. É importante ressaltar que grafos que abrangem duas ou mais disciplinas contêm pesquisadores que atuam em qualquer uma dessas disciplinas ou em combinações das mesmas.

Primeiramente, vamos nos concentrar nas características dos três grafos exibidos na Figura 6. Uma das diferenças mais marcantes entre esses está nas suas densidades, ou na relação entre o número de arestas  $|E|$  e o número de nós  $|V|$ . Enquanto o grau médio  $\bar{k}$  do grafo da Bioquímica é 8,78, o da genética é 5,72 e o da CC é somente 3,33. Isso sugere que, na Bioinformática, os pesquisadores de Bioquímica levam significativamente mais das suas colaborações externas para trabalhos do programa que pesquisadores de Genética e principalmente de CC.

No entanto, quando analisamos o perfil das conexões dentro das redes através da assortatividade de grau  $A_k$ , é a rede da Genética que se destaca. A assortatividade [Newman 2003] mede a tendência de nós similares se conectarem, ou seja, a assortatividade de grau  $A_k$  verifica se nós de grau alto tendem a se conectar com outros nós de grau alto (rede assortativa) ou com nós de grau pequeno (rede disassortativa). O seu valor varia entre  $-1$ , quando a rede é totalmente disassortativa, e  $1$ , quando é totalmente assortativa. Note que enquanto não há nenhuma tendência das redes da Bioquímica e da

CC, na Genética há uma leve tendência de que nós do mesmo grau colaborem entre si. Isso indica que pesquisadores seniores (juniores), que geralmente têm um maior (menor) número de conexões, tendem a se conectar com outros pesquisadores seniores (juniores).

Outro aspecto interessante de se analisar em redes complexas é a distância entre os nós, que pode fornecer percepções valiosas sobre a dinâmica das colaborações dentro da rede. Quanto menor a distância entre dois nós na rede, mais provável é que esses nós venham a interagir no futuro. Assim, note que a distância mínima média  $avg(d_{i,j})$  entre dois nós das redes varia significativamente. Enquanto na Bioquímica  $avg(d_{i,j})$  é 3,55 saltos, na genética é 4,29 e na CC é 7,40, mais que o dobro da Bioquímica. Isso mais uma vez corrobora com a conjectura que os pesquisadores de CC associados à programas de Bioinformática não levam consigo muitas das suas colaborações externas, pois dentro do programa fazem mais pesquisa aplicada, o que geralmente não requer várias especialistas de diferentes linhas de CC. Note também que o diâmetro  $max(d_{i,j})$  das redes são proporcionais à distância mínima média  $avg(d_{i,j})$  das mesmas.

Além dos grafos individuais, é importante analisar os grafos formados a partir de combinações de áreas. Note, por exemplo, que o grafo formado pelos pesquisadores que atuam em Bioquímica e/ou Genética é o mais denso dos analisados, com grau médio  $\bar{k} = 17,50$ . Isso, aliado ao alto coeficiente de aglomeração  $cc = 0,55$ , que também é o mais alto entre os grafos, mostra o alto grau de interação existente entre os pesquisadores de Bioquímica e Genética nos programas de Bioinformática. Para corroborar ainda mais com essa constatação, observe a assortatividade de área  $A_{area}$  dos grafos, que mede o grau de interação entre pesquisadores de áreas diferentes. Quanto maior for  $A_{area}$ , menor é a interação entre pesquisadores de áreas diferentes. Note que  $A_{area}$  para o grafo formado pelos pesquisadores que atuam em Bioquímica e/ou Genética é significativamente menor que aqueles compostos por CC e Bioquímica ou Genética. Mais uma vez, isso sugere que os cientistas da Computação têm tido um papel bem distinto nos trabalhos de bioinformática quando comparados aos cientistas de Bioquímica e Genética.

## 5. Conclusões

Neste trabalho, caracterizamos o conjunto de pesquisadores dos programas de pós-graduação em Bioinformática do Brasil. Primeiro, analisamos a composição em termos de áreas de atuações dos docentes. Concluimos que existe grande diversidade na composição, mas há uma predominância das áreas de CC, Genética e Bioquímica, que respondem juntas por 50% das menções à áreas. Discutimos diferenças importantes entre as principais métricas de produtividade nessas áreas, tendo em vista os pesquisadores de Bioinformática. Quantificamos pela primeira vez no Brasil as diferenças entre o número de publicações em periódicos e conferências. Caracterizamos também redes de coautoría entre pesquisadores dos diferentes programas e áreas. Concluimos que há ainda um percentual baixo de colaborações inter-programas. Com relação às colaborações internas, elas estão em crescimento na UFMG e UFPR, em queda na USP e estáveis no programa da Fiocruz. Vimos ainda que os grafos de colaboração entre as três áreas analisadas são significativamente diferentes e que os pesquisadores de CC estabelecem poucas colaborações no contexto dos programas de Bioinformática se comparados aos de Bioquímica e Genética. Constatamos que os pesquisadores de Bioquímica levam muitos colaboradores externos a participarem dos trabalhos no programa pela densidade da rede de vizinhança formada. Finalmente, notamos que os pesquisadores de Genética têm

tendência a preferirem colaborações com pesquisadores parecidos entre si (em termos de números de colaboradores e áreas de atuação). Esse comportamento não se verificou nas outras áreas. Há ainda muito que se estudar rumo à compreensão da composição e da colaboração de áreas em programas interdisciplinares. Este trabalho é um primeiro passo rumo a esse objetivo no cenário brasileiro.

## Referências

- Balancieri, R., Bovo, A., Medina, V., Pacheco, R., and Barcia, R. (2005). A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na plataforma lattes. *Ciência da Informação*, 34:64–77.
- Bansard, J. Y., Rebholz-Schuhmann, D., Cameron, G., Clark, D., Van Mulligen, E., Beltrame, F., Barbolla, E., Martin-Sanchez, F., Milanesi, L., Tollis, I., Van der Lei, J., and Coatrieux, J. L. (2007). Medical informatics and bioinformatics: A bibliometric study. *IEEE Transactions on Information Technology in Biomedicine*, 11(3):237–243.
- Bollen, J., Rodriguez, M. A., and de Sompel, H. V. (2006). Journal status. *CoRR*, abs/cs/0601030.
- Cainelli, G., de Felice, A., Lamonarca, M., and Zoboli, R. (2006). The publications of italian economists in econlit. quantitative assessment and implications for research evaluation. *Economia Politica*, 0(3):385–423.
- Digiampietri, L., Mena-Chalco, J., Silva, G. S., Oliveira, L., Malheiro, A., and Meira, D. (2012). Dinâmica das relações de coautoria nos programas de pós-graduação em computação no brasil. In *CSBC 2012 - BraSNAM* ().
- Duffy, R. D., Jadidian, A., Webster, G. D., and Sandell, K. J. (2011). The research productivity of academic psychologists: assessment, trends, and best practice recommendations. *Scientometrics*, 89(1):207–227.
- Franceschet, M. (2011). Collaboration in computer science: A network science approach. *Journal of the American Society for Information Science and Technology*, 62(10):1992–2012.
- Freire, V. and Figueiredo, D. (2011). Ranking in collaboration networks using a group based metric. *Journal of the Brazilian Computer Society*, pages 1–12.
- Garfield, E. (1955). Citation indexes for science. a new dimension in documentation through association of ideas. *Science*, 122:1123–1127.
- Garfield, E. (2006). Citation indexes for science. A new dimension in documentation through association of ideas. *Int. J. Epidemiol.*, 35(5):1123–1127.
- Glänzel, W., Janssens, F., and Thijs, B. (2009). A comparative analysis of publication activity and citation impact based on the core literature in bioinformatics. *Scientometrics*, 79:109–129.
- Guan, J. and Gao, X. (2008). Comparison and evaluation of chinese research performance in the field of bioinformatics. *Scientometrics*, 75:357–379.
- Guimaraes, P. (2001). The state of portuguese research in economics: an analysis based on publications in international journals. Working Papers 15, Núcleo de Investigação em Microeconomia Aplicada (NIMA), Universidade do Minho.

- Hayat, Z. and Lyons, K. (2010). The evolution of the cascon community: a social network analysis. In *Proceedings of the 2010 Conference of the Center for Advanced Studies on Collaborative Research, CASCON '10*, pages 1–12, Riverton, NJ, USA. IBM Corp.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572.
- Hogeweg, P. (2011). The roots of Bioinformatics in theoretical biology. *PLoS Computational Biology*, 7(3):e1002021.
- Horn, D. B., Finholt, T. A., Birnholtz, J. P., Motwani, D., and Jayaraman, S. (2004). Six degrees of jonathan grudin: a social network analysis of the evolution and impact of cscw research. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work, CSCW '04*, pages 582–591, New York, NY, USA. ACM.
- Janssens, F., Glänzel, W., and De Moor, B. (2007). Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '07*, pages 360–369, New York, NY, USA. ACM.
- Jeong, S., Lee, S., and Kim, H.-G. (2009). Are you an invited speaker? a bibliometric analysis of elite groups for scholarly events in bioinformatics. *J. Am. Soc. Inf. Sci. Technol.*, 60(6):1118–1131.
- Laender, A., de Lucena, C., Maldonado, J., de Souza e Silva, E., and Ziviani, N. (2008). Assessing the research and education quality of the top Brazilian Computer Science graduate programs. *ACM SIGCSE Bulletin*, 40(2):135–145.
- Martins, W. S., Gonçalves, M. A., Laender, A. H. F., and Ziviani, N. (2010). Assessing the quality of scientific conferences based on bibliographic citations. *Scientometrics*, 83(1):133–155.
- Mena-Chalco, J., Digiampietri, L. A., and Oliveira, L. B. (2012). Perfil de produção bibliográfica dos programas brasileiros de pós-graduação em ciência da computação. In *In: 3º Encontro Brasileiro de Bibliometria e Cientometria (EBBC 2012)*.
- Mena-Chalco, J. P. and Cesar-Jr., R. M. (2009). scriptLattes: An open-source knowledge extraction system from the lattes platform. *Journal of the Brazilian Computer Society*, 15(4):31–39.
- Menezes, G. V., Ziviani, N., Laender, A. H. F., and Almeida, V. (2009). A geographical analysis of knowledge production. In *in Computer Science In Proceedings of the 18th international conference on World Wide Web*, pages 1041–1050.
- Molatudi, M., Molotja, N., and Pouris, A. (2009). A bibliometric study of bioinformatics research in south africa. *Scientometrics*, 81:47–59.
- Newman, M. (2003). The structure and function of complex networks.
- Rodríguez, D. (2006). Publishing performance of spanish academics: 1970-2004. *Spanish Economic Review*, 8(4):255–270.
- Song, M. and Kim, S. (2012). Detecting the knowledge structure of bioinformatics by mining full-text collections. *Scientometrics*, pages 1–19.