CURRICULUM VITÆ

Sophie SCHBATH

Institut National de la Recherche Agronomique

BIOGRAPHIE

Née le 19 décembre 1969 à Nantes, France.
Mariée, 2 enfants.
E-mail : `Sophie.Schbath@inrae.fr`

DIPLOMES

2003 Habilitation à Diriger des Recherches, Université d'Evry.
Deux approches mathématiques de l'analyse des génomes : statistiques des comptages de mots et prédictions en cartographie physique.

1995 Thèse de Mathématiques, spécialité Statistique, Université Paris V.
Étude asymptotique du nombre d'occurrences d'un mot dans une chaîne de Markov et application à la recherche de mots de fréquence exceptionnelle dans les séquences d'ADN, mention très honorable avec les félicitations.

1992 DEA "Modélisation Stochastique et Statistique", Université Paris XI.
Mention assez bien.

1991 Maîtrise de Mathématiques Appliquées aux Sciences Fondamentales, Université Paris XI. Mention assez bien.

1989 DEUG A, Université Paris XI. Mention bien.

1987 Baccalauréat C, Les Ulis (91). Mention assez bien.

EXPERIENCE PROFESSIONNELLE

2018 – Directrice de Recherche 1ère classe, Unité Mathématiques et Informatique Appliquées, du Génome à l'Environnement (MaIAGE), INRAE, Jouy-en-Josas.

2006 – 2018 Directrice de Recherche 2e classe, Unité Mathématique, Informatique & Génome (MIG) puis MaIAGE, INRA, Jouy-en-Josas.

2000 – 2005 Chargée de Recherche 1ère classe, Unité Mathématique, Informatique & Génome (MIG), INRA, Jouy-en-Josas.

1996 – 1999 Chargée de Recherche 2e classe, Unité de Biométrie, INRA, Jouy-en-Josas.

1996 Post-doctorante, Mathematics Department, University of Southern California, Los Angeles, USA.

1992 – 1996 Attachée Scientifique Contractuelle, Unité de Biométrie, INRA, Jouy-en-Josas.

ADMINISTRATION et ANIMATION de la RECHERCHE

Directrice de l'unité *Mathématiques et Informatique Appliquées, du Génome à l'Environnement* (2015 – 2022).

Directrice de l'unité *Mathématique, Informatique et Génome* (2012 – 2014).

Responsable scientifique de la plateforme de bioinformatique Migale (2016 –).

Co-fondatrice et co-directrice du GdR CNRS 3003 *BioInformatique Moléculaire* (BiM) (2006–2009, 2010–2013). `http://www.gdr-bim.u-psud.fr/`

Présidente de la Société Française de BioInformatique (SFBI) (2010 – 2016).

Animatrice du réseau *Statistics for Systems Biology* (anciennement "Statistiques des Séquences Biologiques") (1995 – 2012). `http://www.ssbgroup.fr`

Participation à des instances collectives :
- Commission Administrative Paritaire Locale des Techniciens de la Recherche, INRA Jouy-en-Josas (2014 –).
- Conseil Académique de l'Université Paris-Saclay (2015 – 2019).
- Commission Locale de Développement Durable (2017 –), INRAE Jouy-en-Josas.
- Comité d'Orientation Stratégique RSE d'INRAE (2022 –).

EXPERTISES et CONSEILS SCIENTIFIQUES

Relectrice de 56 articles de journaux depuis 1996 pour
- *Annals of Applied Probability*, *Journal of Applied Probability*, *ESAIM : Probability and Statistics*, *Combinatorics, Probability and Computing*, *Methodology and Computing in Applied Probability*, *Discrete Applied Mathematics*, *Glasnik Matematicki*, *Annales de l'Institut Henri Poincarré*, *Annals of the Institute of Statistical Mathematics*, *Biometrics*, *Statistical Applications in Genetics and Molecular Biology*,
- *Journal of Computational Biology*, *Journal of Mathematical Biology*, *Bioinformatics*, *BMC Bioinformatics*, *IEEE Transactions on Computational Biology and Bioinformatics*, *Journal of Bioinformatics and Computational Biology*, *Journal of Computer & Chemistry*, *INFORMS-Journal of Computing*, *Europhysics Letters*,
- *Nucleic Acid Research*, *Genomics*, *Journal of Molecular Evoluvion*, *Canadian Journal of Microbiology*, *Archae*.

Membre de comités éditoriaux :
- *Scandinavian Journal of Statistics* (sept 2012 - août 2019)
- *Journal of Computational Biology* (2014 - 2019)
- *Methodology and Computing in Applied Probability* (2017 - 2021)
- *Nucleic Acid Research Genomics and Bioinformatics* (avril 2019 -)

Membre de comités de programme de conférences :
- *3rd Annual International Conference on Computational Molecular Biology* (RECOMB), Lyon,

France. Avril 1999.
- *2e Journées Ouvertes : Biologie, Informatique, Mathématique* (JOBIM), Toulouse, France. Mai 2001.
- *3rd Workshop on Algorithms in BioInformatics* (WABI), Budapest, Hongrie. Septembre 2003.
- *8th Annual International Conference on Computational Molecular Biology* (RECOMB), San Diego, Californie, USA. Mars 2004.
- *6e Journées Ouvertes : Biologie, Informatique, Mathématique* (JOBIM), Lyon, France. Juillet 2005.
- *5th Workshop on Algorithms in BioInformatics* (WABI), Eivissa, Espagne. Octobre 2005.
- *7e Journées Ouvertes : Biologie, Informatique, Mathématique* (JOBIM), Bordeaux, France. Juillet 2006.
- *11th Annual International Conference on Computational Molecular Biology* (RECOMB), Oakland, Californie, USA. April 2007.
- *10e Journées Ouvertes : Biologie, Informatique, Mathématique* (JOBIM), Nantes, France. Juin 2009.
- *42e Journées de Statistiques* (JDS), Marseille, France. Mai 2010.
- *12e Journées Ouvertes : Biologie, Informatique, Mathématique* (JOBIM), Paris, France. Juin 2011.
- *13e Journées Ouvertes : Biologie, Informatique, Mathématique* (JOBIM), Rennes, France. Juillet 2012.
- *14e Journées Ouvertes : Biologie, Informatique, Mathématique* (JOBIM), Toulouse, France. Juillet 2013.
- *16e Journées Ouvertes : Biologie, Informatique, Mathématique* (JOBIM), Clermont-Ferrand, France. Juillet 2015.
- *8th International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS 2017), Porto, Portugal. Février 2017.
- *19e Journées Ouvertes : Biologie, Informatique, Mathématique* (JOBIM), Marseille, France. Juillet 2018.
- *20e Journées Ouvertes : Biologie, Informatique, Mathématique* (JOBIM), Nantes, France. Juillet 2019.
- *21e Journées Ouvertes : Biologie, Informatique, Mathématique* (JOBIM), Montpellier, France. Juillet 2020.
- *2ée Journées Ouvertes : Biologie, Informatique, Mathématique* (JOBIM), Paris, France. Juillet 2021. Présidente.

Membre de Conseils Scientifiques :
- *Programme Bioinformatique* français inter-EPST (2000 – 2003).
- Action Concertée Incitative *IMPBio* (2003 - 2004).
- Département *Mathématiques et Informatique Appliquées* de l'INRA (2002 – 2011).
- *Cellule Bioinformatique* de l'INRA (2009 – 2016).
- Cellule de coordination du Métaprogramme *Métaomiques des Ecosystèmes Microbiens* de l'INRA (2010 – 2018).
- Science Advisory Board du projet européen FP7 RADIANT (2013 – 2015)
- Science Advisory Board de l'Institut de Biologie Computationelle de Montpellier (2014 – 2017)
- GdR CNRS 3003 *BioInformatique Moléculaire* (2014 – 2020)
- Conseil du département de Mathématique de l'Université Paris-Saclay (2016 – 2020)
- Conseil de la Graduate School de Mathématiques de l'Université Paris-Saclay (2020 –) ; représentante INRAE.
- Comité de pilotage de l'Institut Convergence *INCEPTION* (2017 –) ; représentante INRAE.
- Comité des programmes de l'Institut Convergence *DataIA* (2019 –) ; représentante INRAE.

Membre d'Instances d'évaluation :

- Commission Scientifique Spécialisée *Mathématique, Bioinformatique et Intelligence Artificielle* de l'INRA (2002 – 2010).
- Comité d'évaluation AERES du laboratoire TIMC de Grenoble (avril 2010).
- Comité d'évaluation scientifique des personnels (COMESP) de l'Institut Pasteur (2016 – 2019).
- Commission Scientifique Spécialisée *Soutien et Pilotage de la Recherche* d'INRAE (2021 – 2024).
- Comité d'évaluation HCERES de l'unité LITIS de Rouen (octobre 2021).

Participations à des jurys de concours :
- Assistant Ingénieur à l'INRA (2008),
- Ingénieur d'Etude à l'INRA (1999, 2019-présidente),
- Ingénieur de Recherche à l'INRA (2009, 2014),
- Ingénieur de Recherche (CIPP) à INRAE (2020-présidente),
- Ingénieur de Recherche Hors Classe (2017-2019, 2021),
- Maître de Conférence à l'INA-PG (2005),
- Maître de Conférence à l'Université Paris-Sud (2011), Rouen (2012), Marseille (2014), Paris-Pierre et Marie Curie (2016), Le Mans (2017), Paris-Diderot (2018),
- Professeur à l'Université de Montpellier 2 (2009), Paris-Diderot (2015, 2018), Université Libre de Bruxelles (2016),
- Chargé de Recherche à l'INRA (2001, 2002, 2013, 2018, 2021),
- Chargé de Recherche 1ère classe à l'INRA (2014, 2016),
- Group leader à l'Institut Pasteur (2015),
- Directeur de Recherche à l'INRA (2007, 2014, 2016, 2018, 2021), à l'INRIA (2011).

Participation à des jurys de thèse :
- Élodie Nédélec, Université Paris XI, examinatrice (2004).
- Gaëlle Gusto, Université Paris XI, encadrante (2004).
- Christelle Melo de Lima, Université de Lyon, rapporteur (2005).
- Narjiss Touyar, université de Rouen, co-encadrante (2006).
- Leonor Palmeira, Université de Lyon, rapporteur (2007).
- Etienne Roquain, Université Paris XI, encadrante (2007).
- Fabrice Touzain, Université de Nancy, présidente (2007).
- Fanny Villers, Université Paris XI, présidente (2007).
- Aude Liefooghe, Université de Lille, membre (2008).
- Simona Grusea, Université de Marseille, rapporteur (2008).
- Elisabeth Ford, Oxford University, rapporteur (2009).
- Lisbeth Carstensen, Copenhagen University, rapporteur (2010).
- Hugo Devillers, Université d'Evry, directrice (2011).
- Jean-Baka Domelevo-Entfellner, Université de Montpellier, rapporteur (2011).
- Abdelkader Behdenna, Université Paris Pierre et Marie Curie, examinatrice (2016).
- Ibrahim Sultan, Université Paris-Saclay, co-encadrante (2019).
- Christophe Menichelli, Université de Montpellier, examinatrice (2019).
- Guillaume Gautreau, Université d'Evry, examinatrice (2020).
- Romain Ménégaux, Université de recherche Paris Sciences et Lettres, examinatrice (2021).

Participation à des jurys de HDR :
- Valentina Boeva, Université Paris 6, examinatrice (2014).
- Pierre Peterlongo, Université de Rennes, rapporteure (2016).
- Guillem Rigaill, Université Paris-Saclay, rapporteure (2020).
- Rayan Chikhi, Sorbonne Université, rapporteure (2021).

VALORISATION et VULGARISATION de la RECHERCHE, ENSEIGNEMENT

Co-responsable du cycle de formation “Bioinformatique par la pratique” de la plate-forme MIGALE (`http://migale.jouy.inra.fr/formations`) depuis 2004.

Activités d’enseignement sur le thème de méthodes statistiques pour la cartographie et l’analyse des génomes ou sur le langage R : en moyenne 45h/an sur 2000-2011, environ 24h/an depuis 2012. Par exemple : *Chaînes de Markov et modèles de Markov cachés pour l’analyse de séquences* (1 jour) dans le Master 2 de BioInformatique de Paris-Diderot, ou *Initiation au langage R* (2 jours) dans le cycle de formation continue de la plateforme Migale.

Participation à la création et à l’animation d’un atelier grand public sur l’alignement de séquences lors des journées portes ouvertes du centre Inra de Jouy-en-Josas à l’occasion des 70 ans de l’Inra (2016).

Exposé de vulgarisation dans le cadre de la K’fêt des Sciences du collège de Bures-sur-Yvette (9 mars 2018) `http://sciencescollegebures.blogspot.fr`

Interview dans le cadre des Tables Ouvertes en BioInformatiques de l’association des Jeunes Bioinformatitiens Français (2016).
`https://jebif.fr/fr/evenements/tobi-tables-ouvertes-en-bioinfo/`

Co-auteur (avec F. Gélis, puis A. Bouvier, puis M. Hoebeke) du logiciel *R’MES* dedié à la Recherche de Mots Exceptionnels dans les Séquences (distribué gratuitement sur le site `http://migale.jouy.inra.fr/outils/mig/rmes`

ORGANISATION DE CONFERENCES

Co-organisation :
- 11e *European Young Statisticians Meeting*, Marly-le-Roi, France. Août 1999.
- 3e *Annual International Conference on Computational Molecular Biology* (RECOMB), Lyon, France. Avril 1999.
- colloque *Mathematics for Biological Networks*, Paris, France (`http://stat.genopole.cnrs.fr/MBN2007/`). Décembre 2007.
- Journées du *GdR Bioinformatique Moléculaire*, Paris, France (`http://www.gdr-bim.u-psud.fr/journees-gdr.php`). Novembre 2009.
- 1er satellite *Bioinformatics for Regulatory Genomics* (BioReg SIG) de ISMB, Boston, USA (`http://light.ece.ohio.edu/bioreg/2010/`). Juillet 2010.
- Ecole-Chercheurs *Métagénomique des Ecosystèmes Microbiens*, Paris, France. Février 2011.
- 2e satellite *Bioinformatics for Regulatory Genomics* (BioReg SIG) de ISMB, Vienne, Autriche (`http://light.ece.ohio.edu/bioreg/2011/`). Juillet 2011.
- Journées du *GdR Bioinformatique Moléculaire*, Paris, France (`http://mig.jouy.inra.fr/?q=fr/journees-gdrbim-2012`). Janvier 2012.
- Journées du *GdR Bioinformatique Moléculaire*, Paris, France (`http://www.gdr-bim.cnrs.fr/coll-2013/`). Novembre 2013.
- 1er workshop *Recent Computational Advances in Metagenomics* (RCAM), en marge de ECCB’14, Strasbourg, France. Septembre 2014.
- 13e *European Conference on Computational Biology* (ECCB), Strasbourg, France (`http://www.`

eccb14.org). Septembre 2014.
- 2e workshop *Recent Computational Advances in Metagenomics* (RCAM), Paris, France. Septembre 2015.
- 3e workshop *Recent Computational Advances in Metagenomics* (RCAM), en marge de ECCB'16, La Haye, Pays-Bas. Septembre 2016.
- 4e workshop *Recent Computational Advances in Metagenomics* (RCAM), Paris, France. Septembre 2017.
- 50e Journées de Statistique, Palaiseau, France. Mai 2018.
- 5e workshop *Recent Computational Advances in Metagenomics* (RCAM), en marge de ECCB'18, Athènes, Grèce. Septembre 2018.
- Assemblée Générale du département MIA de l'Inra, Massy-Jouy, France. Mai 2019.
- 6e workshop *Recent Computational Advances in Metagenomics* (RCAM), Paris, France. Septembre 2019.

Organisation de sessions :
- "Phylogénie" et "Analyse statistique des séquences" aux *Journées MAS : Modélisation pour les Sciences du Vivant*, Grenoble, France. Septembre 2002.
- "Probabilistic problems on words in computational biology" à l'*International Workshop in Applied Probability*, Le Pirée, Grèce. Mars 2004.
- "Probability and Statistics applied to Computational Biology" à l'*International Workshop in Applied Probability*, University of Connecticut, USA. Mai 2006.
- "Applied probability methodology in computational biology" à la 31ème *Conference on Stochastic Processes and their Applications* Paris, France. Juillet 2006.
- "Probability and Statistics applied to Computational Biology" à l'*International Workshop in Applied Probability*, Université de Compiègne, France. Juillet 2008.
- "Probability and Statistics for Genomics" à l'*International Workshop in Applied Probability*, Madrid, Espagne. Juillet 2010.
- " Applied Probability in Computational Biology" à l'*International Workshop in Applied Probability*, Antalya, Turquie. Juin 2014.

PUBLICATIONS

Articles dans revues à comité de lecture

[1] SCHBATH, S. (1995). Compound Poisson approximation of word counts in DNA sequences. *ESAIM : Probability and Statistics*. **1** 1–16.

[2] SCHBATH, S., PRUM, B. and TURCKHEIM, É. DE (1995). Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J. Comp. Biol.* **2** 417–437.

[3] SCHBATH, S. (1997). Coverage processes in physical mapping by anchoring random clones. *J. Comp. Biol.* **4** 61–82.

[4] SCHBATH, S. (1997). An efficient statistic to detect over- and under-represented words in DNA sequences. *J. Comp. Biol.* **4** 189–192.

[5] REINERT, G. and SCHBATH, S. (1998). Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. *J. Comp. Biol.* **5** 223–254.

[6] EL KAROUI, M., BIAUDET, V., SCHBATH, S. and GRUSS, A. (1999). Characteristics of Chi distribution on several bacterial genomes. *Research in Microbiology*. **150** 579–587.

[7] REINERT, G., SCHBATH, S. and WATERMAN, M. (2000). Probabilistic and statistical properties of words : an Overview. *J. Comp. Biol.* **7** 1–46.

[8] SCHBATH, S., BOSSARD, N. and TAVARÉ, S. (2000). The effect of non-homogeneous clone length distribution on the progress of an STS mapping project. *J. Comp. Biol.* **7** 47–58.

[9] SCHBATH, S. (2000). An overview on the distribution of word counts in Markov chains. *J. Comp. Biol.* **7** 193–201.

[10] ROBIN, S. and SCHBATH, S. (2001). Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comp. Biol.* **8** 349–359.

[11] ROBIN, S., DAUDIN, J.-J., RICHARD, H., SAGOT, M.-F. and SCHBATH, S. (2002). Occurrence probability of structured motifs in random sequences. *J. Comp. Biol.* **9** 761–773.

[12] SCHBATH, S. (2003). Statistical methods in physical mapping. *Encyclopedia of the Human Genome*, 434, Nature Publishing Group. (`http://www.ehgonline.net/mathematical.asp`)

[13] SCHBATH, S. (2004). A la recherche de mots de fréquence exceptionnelle dans les génomes. *Images des Mathématiques*.

[14] GUSTO, G. and SCHBATH, S. (2005). FADO : a statistical method to detect favored or avoided distances between motif occurrences using the hawkes' model. *Statistical Applications in Genetics and Molecular Biology*. **4**, Article 24.

[15] MATIAS, C., SCHBATH, S., BIRMELÉ, E., DAUDIN, J.-J. and ROBIN, S. (2006). Network motifs : mean and variance for the count. *REVSTAT*. **4** 31–51.

[16] STEFANOV, V., ROBIN, S., and SCHBATH, S. (2007). Waiting times for clumps of patterns and for structured motifs in random sequences. *Discrete Applied Mathematics*. **155**, 868–880.

[17] ROQUAIN, E. and SCHBATH, S. (2007). Improved compound Poisson approximation for the number of occurrences of multiple words in a stationary Markov chain. *Adv. Appl. Prob.* **39** 1–13.

[18] ROBIN, S., SCHBATH, S. and VANDEWALLE, V. (2007). Statistical tests to compare motif count exceptionalities. *BMC Bioinformatics* **8 :84**, 20 pages.

[19] HALPERN, D., CHIAPELLO, H., SCHBATH, S., ROBIN, S., HENNEQUET-ANTIER, C., GRUSS, A. and EL KAROUI, M. (2007). Identification of DNA motifs implicated in maintenance of bacterial core genomes by predictive modelling. *PLoS Genetics*. **3(9)** e153.

[20] TOUZAIN, F., SCHBATH, S., DEBLED-RENNESSON, I., AIGLE, B., LEBLOND, P. and KUCHEROV, G. (2008). SIGffRid : a tool to search for σ factor binding sites in bacterial genomes using comparative approach and biologically driven statistics. *BMC Bioinformatics*. **9 :73** 1–23.

[21] PICARD, F., DAUDIN, J.-J., KOSKAS, M., SCHBATH, S. and ROBIN, S. (2008). Assessing the exceptionality of network motifs. *J. Comp. Biol.* **15 :1** 1–20.

[22] TOUYAR, N., SCHBATH, S., CELLIER, D. and DAUCHEL, H. (2008). Poisson approximation for the number of repeats in a Markov chain model. *J. Appl. Prob.* **45** 440–455.

[23] MERCIER, R., PETIT, M.-A., SCHBATH, S., ROBIN, S., EL KAROUI, M., BOCCARD, F. and ESPELI, O. (2008). The MatP/matS site specific system organizes the Terminus region of the E. coli chromosome into a Macrodomain. *Cell*. **135** 475–485.

[24] SCHBATH, S., LACROIX, V. and SAGOT, M.-F. (2009). Assessing the exceptionality of coloured motifs in networks. *EURASIP Journal on Bioinformatics and Systems Biology*. **ID 616234** 1–9.

[25] REYNAUD-BOURET, P. and SCHBATH, S. (2010). Adaptive estimation for Hawkes' processes ; Application to genome analysis. *Annals of Statistics*. **38 (5)** 2781–2822.

[26] TOUZAIN, F., PETIT, M.-A., SCHBATH, S. and EL KAROUI, M. (2011). DNA motifs that sculpt the bacterial chromosome. *Nature Reviews Microbiology*. **9** 15–26.

[27] STEFANOV, V., ROBIN, S. and SCHBATH, S. (2011). Occurrence of structured motifs in random sequences : Arbitrary number of boxes. *Discrete Applied Mathematics*. doi :10.1016/j.dam.2010.12.023.

[28] DEVILLERS, H., CHIAPELLO, H., SCHBATH, S. and EL KAROUI, M. (2011). Robustness assessment of whole bacterial genome segmentations. *Journal of Computational Biology*. **18** 1155–1165.

[29] DEVILLERS, H. and SCHBATH, S. (2012). Separating significant matches from spurious matches in DNA sequences. *Journal of Computational Biology*. **19** 1–12.

[30] FAYYAZ, A., LAUNAY, G., SCHBATH, S., GIBRAT, J.-F. and RODOLPHE, F. (2012). Statistical significance of threading scores. *Journal of Computational Biology*. **19** 13–29.

[31] SCHBATH, S., MARTIN, V., ZYTNICKI, M., FAYOLLE, J., LOUX, V. and GIBRAT, J.-F. (2012). Mapping reads on a genomic sequence : an algorithmic overview and a practical comparative analysis. *Journal of Computational Biology*. **19** 796–813.

[32] DE PAEPE, M., HUTINET, G., SON, O., AMARIR-BOUHRAM, J., SCHBATH, S. and PETIT, M.-A. (2014). Temperate phages acquire DNA from defective prophages by relaxed homologous recombination : The role of Rad52-like recombinases. *PLOS Genetics*. **10(3)** e1004181.

[33] MASSIP, F., SHEINMAN, M., SCHBATH, S. and ARNDT, P. (2015). How evolution of genomes is reflected in exact DNA sequence match statistics. *Molecular biology and evolution*. **32** 524–535.

[34] MASSIP, F., SHEINMAN, M., SCHBATH, S. and ARNDT, P. (2016). Comparing the statistical fate of paralogous and orthologous sequences. *Genetics*. **204** 1–7.

[35] BENOIT, G., PETERLONGO, P., MARIADASSOU, M., DREZEN, E., SCHBATH, S., LAVENIER, D. and LEMAITRE, C. (2016). Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Computer Science*. **2** e94.

[36] BENOIT, G., MARIADASSOU, M., ROBIN, S., SCHBATH, S., PETERLONGO, P. and LEMAITRE, C. (2020). Simkamin : fast and resource frugal de novo comparative metagenomics. *Bioinformatics*. **36**.

[37] HUREL, J., SCHBATH, S., BOUGEARD, S., ROLLAND, M., PETRILLO, M. and F., T. (2020). Dugmo : tool for the detection of unknown genetically modified organisms with high-throughput sequencing data for pure bacterial samples. *BMC Bioinformatics*. **21**.

[38] SULTAN, I., FROMION, V., SCHBATH, S. and NICOLAS, P. (2020). Statistical modelling of bacterial promoter sequences for regulatory motif discovery with the help of transcriptome data : application to Listeria monocytogenes. *Journal of the Royal Society Interface*. **17**.

[39] AUBERT, J., SCHBATH, S. and ROBIN, S. (2021). Model-based biclustering for overdispersed count data with application in microbial ecology. *Methods in Ecology and Evolution*. **12** 1050–1061.

[40] BIZE, A., MIDOUX, C., MARIADASSOU, M., SCHBATH, S., FORTERRE, P. and DA CUNHA, V. (2021). Exploring short k-mer profiles in cells and mobile elements from archaea highlights the major influence of both the ecological niche and evolutionary history. *BMC Genomics*. **22**.

[41] MARIADASSOU, M., NOUVEL, L.-X., CONSTANT, F., MORGAVI, D., RAULT, L., BARBEY, S., HELLOIN, E., RUÉ, O., SCHBATH, S., LAUNAY, F., SANDRA, O., LEFEBVRE, R., LE LOIR, Y., GERMON, P., CITTI, C. and EVEN, S. (2023). Microbiota members from body sites of dairy cows are largely shared within individual hosts throughout lactation but sharing is limited in the herd. *Animal Microbiome*. **5** 32.

Ouvrages / Chapitres d'ouvrages

[42] SCHBATH, S. and BOUVIER, A. (1998). Finding words with unexpected frequencies in DNA sequences. In *Explorapedia of Statistical and Mathematical Techniques for use in Research and Technology* (`http://www.bioss.sari.ac.uk/smart/unix/intro/slides/home.htm`).

[43] REINERT, G. and SCHBATH, S. (1999). Compound Poisson approximations for occurrences of multiple words. Dans *Statistics in Genetics and Molecular Biology*, (F. Seillier, ed.). IMS Lecture Notes-Monograph Series. **33** 257–275.

[44] ROBIN, S., RODOLPHE, F. and SCHBATH, S. (2003). *ADN, mots et modèles*. BELIN.

[45] REINERT, G., SCHBATH, S. and WATERMAN, M. (2005). Statistics on words with applications to biological sequences. Dans *Applied Combinatorics on Words*, (J. Berstel and D. Perrin, ed.). Cambridge University Press.

[46] ROBIN, S., RODOLPHE, F. and SCHBATH, S. (2005). *DNA, Words and Models*. Cambridge University Press.

[47] SCHBATH, S. and ROBIN, R. (2009). How can pattern statistics be useful for DNA motif discovery ? Dans *Scan Statistics – Methods and Applications*, (J. Glaz, I. Pozdnyakov, and S. Wallenstein, eds.). Statistics for Industry and Technology. Birkhauser.

[48] SCHBATH, S. and HOEBEKE, M. (2011). R'MES : a tool to find motifs with a significantly unexpected frequency in biological sequences. Dans *Advances in genomic sequence analysis and pattern discovery*, (L. Elnitski, O. Piontkivska, and L. Welch, eds.). Science, Engineering, and Biology Informatics, vol. 7. World Scientific.

Rapports techniques, Logiciels

[49] GÉLIS, F. and SCHBATH, S. (1996). *R'MES : Recherche de Mots Exceptionnels dans les Séquences d'ADN – Version 1*. Notice d'utilisation. INRA, Biométrie, 78352 Jouy-en-Josas, France.

[50] BOUVIER, A., GÉLIS, F. and SCHBATH, S. (1999). *R'MES : Recherche de Mots Exceptionnels dans les Séquences d'ADN – Version 2*. Guide de l'utilisateur. INRA, Biométrie, F78352 Jouy-en-Josas.

[51] EL KAROUI, M. and SCHBATH, S. (2001). Identification de motifs significativement sur- ou sous-représentés dans un génome : le cas de `gctggtgg` dans le génome d'*Escherichia coli*. Rapport technique pour illustrer un cours de Probabilités de l'Ecole Polytechnique (10 pages).

[52] SCHBATH, S. (2006). Statistics of motifs. Lecture notes for *Atelier INSERM Identification of non-coding functional regions in genome*, La Londes-les-Maures, April 27-28 (10 pages). http://migale.jouy.inra.fr/outils/mig/rmes/atelier-inserm2006.pdf

[53] HOEBEKE, M. and SCHBATH, S. (2006). *R'MES : Finding exceptional motifs, version 3*. User guide. http://migale.jouy.inra.fr/outils/mig/rmes/rmes3.01.userGuide.pdf

[54] SCHAEFFER, B. and SCHBATH, S. (2010, 2016). *Initiation à R*.

Conférences Invitées

[55] SCHBATH, S. (1997). Predicting progress in a physical mapping project by anchoring random clones using coverage processes. Conférence invitée dans *Mathematical Statistics and its Application to Biosciences*. ISI Satelite Meeting, Rostock, Germany. 31 août – 4 septembre.

[56] SCHBATH, S. (2000b). (i) Introduction to the problem of finding words with unexpected frequency in DNA sequences : motivation, word counts, models, periodic structure of words, word count distribution, (ii) Gaussian approximation of the word count distribution and application, (iii) Poisson approximation and the chen-stein method. Dans *Semester in Bioinformatics*. Uppsala, Sweden. 24–25 février.

[57] SCHBATH, S. (2000). Modèles markoviens dans l'analyse statistique des séquences. Conférence invitée dans *Journées TAS, Traitement et Analyse de Séquences*. Evry, France. 22–24 novembre.

[58] SCHBATH, S. (2001). Distribution of word counts in DNA sequences and quality of approximations. Conférence invitée dans *23rd European Meeting of Statisticians*. Funchal, Madeira. 13–18 août.

[59] SCHBATH, S. (2002). Exceptional motifs in biological sequences. Dans *Maps, Sequences and Genomes*. University of Southern California, Los Angeles, USA. 31 mai – 2 juin (résumé).

[60] SCHBATH, S. (2004). Overview on probabilistic problems on words in computational biology. Dans *International Workshop in Applied Probability*. University of Piraeus, Greece. 22-25 mars.

[61] SCHBATH, S. (2004). Modèles statistiques et analyse de génomes. Dans les *36èmes Journées de Statistique*. Montpellier, France. 22-28 mai.

[62] SCHBATH, S. (2004). Modelling the dependence between sequence motifs. Dans *6th World Bernoulli Congress*. Barcelonne, Espagne. 26-30 juillet.

[63] SCHBATH, S. (2005e). Statistical problems arising in physical mapping. Dans *Workshop on Stat. in Genomics and Proteomics*. Estoril, Portugal. 6-8 octobre.

[64] SCHBATH, S. (2005f). The statistical world of motifs on genomes. Dans *SemStat Summer School*. University of Warwick, UK. 11-13 septembre.

[65] SCHBATH, S. (2006b). Statistics of motifs. Dans *Atelier INSERM Identification de régions non codantes fonctionnelles dans les génomes*. La Londes-les-Maures. 27-28 avril.

[66] SCHBATH, S. (2006a). Network motifs : mean and variance for the count. Dans *International Workshop on Applied Probability*. University of Connecticut, USA. 15-18 mai.

[67] SCHBATH, S. (2007). Assessing the exceptionality of network motifs. In *Workshop on Statistics in Genomics and Proteomics*. Centro Internacional de Matemática, Coimbra, Portugal. 9-10 mars

[68] SCHBATH, S. (2008). Occurrences of structured motifs along DNA sequences. In *International Workshop on Applied Probability*. Université Technologique de Compiègne, France. 7-10 juillet.

[69] SCHBATH, S. (2008). The statistical world of motif occurrences along DNA sequences. In *Workshop Hitting, returning and matching in dynamical systems, information theory & mathematical biology*. Eindhoven, The Netherlands. 3-7 novembre.

[70] SCHBATH, S. (2008). Statistical analysis of biological networks ; Assessing the exceptionality of network motifs. In *Approches quantitatives de la complexité biologique*. PRES UniverSud Paris, Orsay, France. 5-6 mai.

[71] SCHBATH, S. (2009). Statistics of biological network motifs ; A compound poisson approximation for their count in random graphs ? In *Progress in Stein's method*. Singapour. 12-16 janvier.

[72] SCHBATH, S. (2009). R'MES : Finding exceptional motifs in sequences. In *Bioinformatics Open Source Conference*. Stockholm, Sweden. 28 juin.

[73] SCHBATH, S. (2010). Statistics of biological network motifs. In *Seminar of the PhD program : Complex systems for postgenomic biology*. Cancer Research Institute, Torino, Italy. 24 février.

[74] SCHBATH, S. (2010). Statistique de mots en génomique : ce qui a été fait, ce qu'il reste à faire. In *Journées de Statistiques du Sud*. Mèze, France. 24 juin.

[75] SCHBATH, S. (2011). Motif-based comparison of biological networks. In *Workshop on Discrete Mathematics and Probability in Networks and Population Biology*. Singapour. 9-13 mai.

[76] SCHBATH, S. (2011). Statistical models and analyses for biological networks. In *Networks research cluster workshop*. Oxford, UK. 2 juin.

[77] SCHBATH, S. (2012). Statistics of network motifs. In *Computational Biology Symposium*. Los Angeles, USA 30 mars - 1er avril.

[78] SCHBATH, S. (2012). Metabolic network comparison based on coloured motif occurrences. In *International Workshop on Applied Probability*. Jerusalem, Israel. 11-14 juin.

[79] SCHBATH, S. (2012b). Separating significant matches from spurious matches in DNA sequences. In *Conférence en l'honneur des 65 ans d'Alain Guénoche*. Marseille, France. October 25-26.

[80] SCHBATH, S. (2013). Mapping reads on a genomic sequence : a practical comparative analysis. In *Kick-of meeting of PF7 RADIANT project*. Manchester, UK. January 14-15.

[81] SCHBATH, S. (2014a). Assessing the enrichment significance of a position weight matrix along a DNA sequence. In *International Workshop in Applied Probability*. Antalya, Turkey. June 16-19 (Invited Lecture).

[82] SCHBATH, S. (2014b). La bioinformatique et sa société savante à l'échelle nationale. In *Rencontres METIC*. Montpellier, France. October 21.

[83] SCHBATH, S. (2015a). A la recherche de motifs statistiquement sur-représentés dans les génomes : des mots aux matrices poids-position. In *Séminaire du CMAP*. École Polytechnique, Palaiseau, France. March, 10.

[84] SCHBATH, S. (2015b). Statistics of motifs : how to deal with position-weight matrices. In *Meeting of FP7 RADIANT project*. Naples, Italy. July 2-3 (invited lecture).

[85] SCHBATH, S. (2017). Une histoire de mots innattendus et de génomes. In *Journées ALEA*. Luminy, France. March 20-24 (invited lecture). Video : `http://library.cirm-math.fr/Record.htm?idlist=1&record=19282373124910005559`.

[86] SCHBATH, S. (2018). The french Microbial Ecosystems and Meta-omics (MEM) metaprogramme from INRA. In *Hellenic Bioinformatics*. Thessalonica, Greece. November 16-18 (invited lecture).

[87] LOUX, V. and SCHBATH, S. (2018). Des microbes dans mon fromage ? In *Journées Nationales de la Science Ouverte*. Paris, France. December 6 (invited lecture).

[88] SCHBATH, S. (2019). Quels microbes pour fabriquer un nouveau jus de lupin fermenté ? Le text-mining à la rescousse ! In *VisaTM Days*. Paris, France. November 15 (invited lecture).

[89] SCHBATH, S. (2019). Text-mining : a complementary approach to bioinformatics for research in microbiology. In *DataIA Days : IA and Agronomics*. Université Paris-Saclay, France. December 4 (invited lecture).

[90] SCHBATH, S. (2021). GreenMaIAGE : initiatives pour des pratiques plus écoresponsables. In *Séminaire de l'unité GABI d'INRAE*. Jouy-en-Josas, France. May 31 (invited lecture).

[91] SCHBATH, S. (2021). GreenMaIAGE : bilan ges et initiatives pour des pratiques plus écoresponsables. In *Journée Impact carbone de la recherche et du numérique du Pôle IMABS*. Toulouse, France. September (invited lecture).

[92] SCHBATH, S. (2021). GreenMaIAGE : bilan ges et initiatives pour des pratiques plus écoresponsables. In *Séminaire du réseau des relais Développement Durable du centre INRAE de Versailles*. Versailles, France. October 22 (invited lecture).

[93] SCHBATH, S. (2021). GreenMaIAGE : bilan ges et initiatives pour des pratiques plus écoresponsables. In *Séminaire du département AgroEcoSystem d'INRAE*. Jouy-en-Josas, France. Novembre 22 (invited lecture).

[94] SCHBATH, S. (2022). MathNum au coeur des enjeux RSE d'INRAE. In *Assemblée Générale du département MathNum d'INRAE*. Clermont-Ferrand, France. May 18 (Invited talk).

[95] SCHBATH, S. (2022). Faire son BGES avec GES-1point5 et après ? retex de l'unité MaIAGE. In *Réunion des directeurs et directrices d'unité du centre INRAE IdF-Versailles*. Versailles, France. September 12 (Invited talk).

[96] SCHBATH, S. (2022). Faire son BGES avec GES-1point5 et après ? retex de l'unité MaIAGE. In *Conseil du centre INRAE IdF-Versailles*. Versailles, France. September 26 (Invited talk).

[97] SCHBATH, S. (2022). MaIAGE engagée dans la réduction de son empreinte carbone. In *Assemblée Générale de l'Unité de recherche HYCAR*. Antony, France. September 22 (Invited talk).

[98] SCHBATH, S. (2022). MaIAGE engagée dans la réduction de son empreinte carbone. In *Atelier participatif Développement Durable et Responsabilité Sociétale de l'unité de recherche GéoSciences*. Rennes, France. November 17 (Invited talk).

[99] SCHBATH, S. (2022). MaIAGE engagée dans la réduction de son empreinte carbone. In *Journées scientifiques de l'IDEEV*. Rennes, France. December 2 (Invited talk).

[100] SCHBATH, S. (2023). MaIAGE engagée dans la réduction de son empreinte carbone : une analyse de la démarche de transition à destination des dus. In *Formation des directeurs et directrices d'unité, GDR Labos-1point5*. Remote, France. February 15 (Invited talk).

[101] SCHBATH, S. (2023). MaIAGE engagée dans la réduction de son empreinte carbone. In *Workshop CO2ERASE, Université Dauphine*. Paris, France. June 22 (Invited talk).

Communications orales (hors posters)

[102] SCHBATH, S. (1994). Identification de motifs exceptionnels par l'étude statistique des comptages de "trains". Dans *Recherche de Motifs dans les Séquences*. GREG et GDR "Informatique et Génomes", Marseille, France. 24-25 février (résumé).

[103] SCHBATH, S. (1994). Étude des comptages de mots dans des séquences d'ADN et approximations par des lois de Poisson. Dans *XXVIèmes Journées de Statistique*. Association pour la Statistique et ses Utilisations, Neuchâtel, Switzerland. 24-27 mai (résumé).

[104] SCHBATH, S. (1994). Recherche de motifs de fréquence exceptionnelle dans les séquences d'ADN. Dans *Forum InterDisciplinaire "Génome et Informatique"*. GREG et GDR "Informatique et Génomes", Aussois, France. 15-17 juin (résumé).

[105] SCHBATH, S. (1995). Statistiques des comptages de mots dans les séquences d'ADN. Dans *XXVIIèmes Journées de Statistique*. Association pour la Statistique et ses Utilisations, Jouy-en-Josas, France. 15-19 mai (résumé).

[106] SCHBATH, S. (1995). Statistics of counts of words in DNA sequences. Dans *21st European Meeting of Statisticians*. Aarhus University, Denmark. 21-25 août (résumé).

[107] SCHBATH, S. (1997). Predicting progress in physical mapping projects without homogeneity assumptions. Dans *Statistics and Inference in Molecular Biology*. Program in Mathematics and Molecular Biology, Santa Fe, USA. 14-19 janvier (résumé).

[108] SCHBATH, S. (1998). Approximation for counts of multiple words in biological sequences. Dans *Workshop on Mathematical and Statistical Aspects of Molecular Biology*. University of Wales College of Medecine, Cardiff, GB. 6-7 avril (résumé).

[109] SCHBATH, S. (1998). How to find exceptional words in biological sequences. Dans *Analyse Structurale et Fonctionnelle d'un génome*. Séminaire Algorithme et Biologie, Institut Pasteur, Paris, France. 8-10 décembre.

[110] SCHBATH, S. (1999). A new method for protein classification based on motifs. Dans *Electrophoresis Forum'99*. Rouen, France. 24-26 novembre (résumé).

[111] SCHBATH, S. (2000). Finding protein interactions by analyzing occurrences of multiple motifs along DNA sequences. Dans *Third Danish-French workshop on spatial statistics and image analysis in biology*. Luminy, France. 7-10 mars.

[112] SCHBATH, S. and GUSTO, G. (2002). Analyse statistique de la corépartition de motifs le long d'une séquence. Dans *Journées MAS : Modélisation pour les Sciences du Vivant*. Grenoble, France. 2-4 septembre (résumé).

[113] SCHBATH, S. (2002b). Statistiques des comptages de mots dans les séquences. Dans *Séminaire Algorithme et Biologie*. Lyon, France. 1-3 octobre (résumé).

[114] BOURGAIT, I., CHIAPELLO, H., HENNEQUET-ANTIER, C., ROBIN, S., SCHBATH, S., GRUSS, A. and EL KAROUI, M. (2003). Genomic distribution of short motifs involved in DNA repair in pathogenic and non pathogenic *E. coli* Dans *Second European Conference on Computational Biology*. Paris, France. 27-30 septembre (**selected short paper**, 7–9).

[115] SCHBATH, S. (2005b). La recherche de mots de fréquence exceptionnelle dans une séquence : état de l'art et derniers résultats. Dans *Séminaire Probabilités, Optimisation, Contrôle*. INRIA Rocquencourt, France. 20 janvier (résumé).

[116] SCHBATH, S. (2005g). À la recherche de motifs exceptionnels dans les génomes. Dans *Séminaire Mathématiques Appliquées*. Université Paris V, France. 15 avril (résumé).

[117] SCHBATH, S. (2005h). À la recherche de motifs exceptionnels dans les génomes : exemples de problèmes statistiques. Dans *Séminaire Modèles Stochastiques*. Ecole Polytechnique, Palaiseau, France. 23 mai (résumé).

[118] TOUZAIN, F., SCHBATH, S., DEBLED-RENNESSON, I., AIGLE, B., LEBLOND. and KUCHEROV, G. (2005). SIGffRid : Programme de recherche des sites de fixation des facteurs de transcription par approche comparative. Dans *Journées Ouvertes Biologie Informatique Mathématiques*. Lyon, France. 6-8 juillet (**selected long paper**, 417–426).

[119] SCHBATH, S. (2005c). Modeling the dependence between sequence motifs. Dans *Journées Algorithmique Génomique*. University of Paris XI, Orsay, France. 24-25 novembre (résumé).

[120] SCHBATH, S. (2005a). Approximation de poisson pour le nombre de répétitions dans une séquence markovienne. Dans *Journées Algorithmique Génomique*. University of Paris XI, Orsay, France. 24-25 novembre (résumé).

[121] SCHBATH, S. (2005d). Network motifs : mean and variance for the count. Dans *CompBioNet*. Lyon, France. 5-7 décembre (texte intégral).

[122] SCHBATH, S. (2006c). Towards exceptional motifs in biological networks. Dans *Workshop Statistical Methods for Post-Genomics data*. INSA, Toulouse, France. 30-31 mars (résumé).

[123] PICARD, F., DAUDIN, J.-J., SCHBATH, S. and ROBIN, S. (2006a). Assessing the exceptionality of network motifs. In *Journée thématique Réseaux d'interactions : analyse, modélisation et simulation*. Université de Lyon, France. Octobre, 29 (article court).

[124] SCHBATH, S. (2006a). Comment tester qu'un motif est significativement plus exceptionnel dans une séquence que dans une autre ? In *Groupe de Travail en Génomique Comparative*. Université de Nantes, France. October, 12-13 (résumé).

[125] SCHBATH, S. (2007b). Assessing the exceptionality of network motifs. In *Tripartite meeting BIOSS/INRA/Biometris*. University of Wageningen, The Netherlands (résumé).

[126] PICARD, F., DAUDIN, J.-J., KOSKAS, M., SCHBATH, S. and ROBIN, S. (2007). Assessing the exceptionalities of network motifs. In *Journées Ouvertes Biologie Informatique Mathématiques (JOBIM)*, (C. Brun and G. Didier, ed.), Marseille, France (**selected long paper**, 235–241).

[127] ROBIN, S., SCHBATH, S. and VANDEWALLE, V. (2007). Statistical tests to compare motif count exceptionalities. In *Journées Ouvertes Biologie Informatique Mathématiques (JOBIM)*, (C. Brun and G. Didier, ed.), Marseille, France (**selected long paper**, 57–62).

[128] DEVILLERS, H., CHIAPELLO, H., EL KAROUI, M. and SCHBATH, S. (2009). How to measure the robustness of bacterial genome comparisons ? In *Journées Ouvertes Biologie Informatique Mathématiques (JOBIM)*, (E. Rivals and I. Rusu, ed.), Nantes, France (**selected long paper**, 25–30).

[129] DEVILLERS, H., CHIAPELLO, H., SCHBATH, S. and EL KAROUI, M. (2010). Assessing the robustness of complete bacterial genome segmentations. In *RECOMB-CG 2010*, (E. Tannier, ed.). Lecture Notes in Bioinformatics. **6398** 173-187 (**selected long paper**).

[130] KOSKAS, M., GRASSEAU, G., BIRMELÉ, E., SCHBATH, S. and ROBIN, S. (2011). Nemo : Fast count of network motifs. In *Journées Ouvertes Biologie Informatique Mathématiques (JOBIM)*, (E. Barillot, C. Froidevaux, and E. Rocha, ed.), Paris, France (**selected long paper**, 53–60).

[131] FAYYAZ, A., LAUNAY, G., SCHBATH, S., GIBRAT, J.-F. and RODOLPHE, F. (2011). How significant is a threading score ? In *Journées Ouvertes Biologie Informatique Mathématiques (JOBIM)*, (E. Barillot, C. Froidevaux, and E. Rocha, ed.), Paris, France (**selected long paper**, 27–34).

[132] KOSKAS, M., GRASSEAU, G., BIRMELÉ, E., SCHBATH, S. and ROBIN, S. (2011). Nemo : Fast count of network motifs. In *Modèles et Analyse de Réseaux : Approches Mathématiques et Informatique (MARAMI)*. Grenoble, France. October 19-21.

[133] SCHBATH, S., MARTIN, V., ZYTNICKI, M. , FAYOLLE, J., LOUX, V. and GIBRAT J.-F. (2012). Mapping reads on a genomic sequence : a practical comparative analysis. In *Journées Ouvertes Biologie Informatique Mathématiques (JOBIM)*, (Coste, F. and Tagu, D., ed.), Rennes, France (**selected long paper**, 183–190).

[134] DUMAZERT, J., STEPHAN, J.-Y., PETIT, M.-A. and SCHBATH, S. (2013). Assessing the enrichment significance of a possition weight matrix (PWM) along a DNA sequence. In *Journées Ouvertes Biologie Informatique Mathématiques (JOBIM)*, (C. Gaspin and e. Lindley, N., ed.), Toulouse, France (**selected long paper**, 25–34).

[135] AUBERT, J., ROBIN, S. and SCHBATH, S. (2014). Metagenomics data analysis using a latent block model : application to plant-microbial communities interactions in the rhizosphere. In *9th European Conference on Mathematical and Theoretical Biology (ECMTB)*. Gothenburg, Sweden. June 15-19 (résumé).

[136] AUBERT, J., SCHBATH, S., MOUGEL, C. and ROBIN, S. (20). Latent block model for ecological abundance data. In *30th European Meeting of Statisticians*. Amsterdam, Nederlands. July 6-10 (résumé).

[137] AUBERT, J., SCHBATH, S. and ROBIN, S. (2016a). Latent block model for metagenomic data. In *17e Journées Ouvertes Biologie Informatique Mathématiques (JOBIM 2016)*. Lyon, France. June 28-30 (**Selected long paper**).

[138] AUBERT, J., SCHBATH, S. and ROBIN, S. (2016b). Latent block model for metagenomic data. In *Workshop on Recent Computational Advances in Metagenomics (RCAM)*. The Hague, Nederlands. September 4 (résumé).

[139] BENOIT, G., PETERLONGO, P., MARIADASSOU, M., DREZEN, E., SCHBATH, S., LAVENIER, D. and LEMAITRE, C. (2017). Simka : large scale de novo comparative metagenomics. In *Journées Ouvertes Biologie Informatique Mathématiques (JOBIM)*, (C. Lhoussaine and Touzet, H., ed.), 3–5, Lille, France. (**Selected long paper**).

[140] SCHBATH, S. (2018). À la recherche de mots exceptionnels dans les génomes. In *K'fêt des sciences du collège de La Guyonnerie*. Bures-sur-Yvette, France. March 9.

[141] SULTAN, I., SCHBATH, S. and NICOLAS, P. (2018). Statistical modelling of bacterial promoter sequences for regulatory motif discovery with the help of transcription profiles. In *50e Journées de Statistique de la SFdS (JdS'2018)*. Palaiseau, France. May 28 - June 1st (résumé).

[142] SULTAN, I., SCHBATH, S. and NICOLAS, P. (2018). Statistical modeling of bacterial promoter sequences for regulatory motif discovery using expression data. In *Journées Ouvertes Biologie, Informatique et Mathématiques (JOBIM)*. Marseille, France. July 3-6 (**Selected long paper**).

[143] GOUTORBE, B., ABRAHAM, A.-L., MARIADASSOU, M., PLAUZOLLES, A., BIDAUT, G., HALFON, P. and SCHBATH, S. (2021). Shallow shotgun metagenomics as a cost-effective and accurate alternative to wgs for taxonomic profiling and clinical diagnosis. In *Journées Ouvertes Biologie, Informatique et Mathématiques (JOBIM)*. Paris, France. July 6-9 (**Selected long paper**).

ENCADREMENT

Thèses

GUSTO, G. (décembre 2004). *Estimation dans un modèle de Hawkes et application à l'étude de la corépartition de motifs le long d'un génome*, Université Paris XI. Co-encadrement avec P. Massart.

TOUYAR, N. (janvier 2006). *Approximation de la loi du nombre de répétitions dans une chaîne de Markov*, Université de Rouen. Co-encadrement avec D. Cellier et H. Dauchel.

ROQUAIN, E. (octobre 2007). *Motifs exceptionnels dans des séquences hétérogènes. Contributions à la théorie et à la méthodologie des tests multiples*, Université Paris XI.

DEVILLERS, H. (février 2011). *Statistiques des comparaisons de génomes complets bactériens*, Université d'Evry. Co-encadrement avec M. El Karoui

MASSIP, F. (octobre 2015). *Développement de modèles d'évolution chromosomique par duplications segmentaires et leurs conséquences sur les scores d'alignement*, Université Paris-Saclay. Co-encadrement avec P. Arndt.

AUBERT, J. (février 2017). *Analyse statistique de données biologiques à haut débit*, Université Paris-Saclay. Co-encadrement avec S. Robin.

SULTAN, I. (juin 2019). *Statistical modeling of bacterial promoter sequences for regulatory motif discovery*, Université Paris-Saclay. Co-encadrement avec P. Nicolas.

GOUTORBE, G. (démarrée en septembre 2019). *Développement et application d'une méthode précise et efficace pour l'analyse du microbiote humain à visée clinique*, Université Paris-Saclay. Co-encadrement avec P. Halfon et G. Bidaut.

Comités de thèses

BOISSIN, A. (décembre 2006). *Prédiction markovienne in silico des régions constantes et variables des lentivirus*, Université Lyon I. Membre du comité de thèse.

CATHERINE ENG (2009). *Utilisation de HMM à la détection du transfert horizontal chez les streptocoques*, Université Nancy. Membre du comité de thèse.

ANA KOZOMARA (2009). *Prédiction de gènes d'ARN dans les séquences génomiques*, Université Toulouse. Membre du comité de thèse.

DAVID ARMISEN (2009). *Caractérisation des éléments structuraux et fonctionnels communs aux gènes orthologues chez les plantes*, Université d'Evry. Membre du comité de thèse.

CERVIN GUYOMAR (2018). *Développement et applications d'outils d'analyse métagénomique des communautés microbiennes associées aux insectes*, Université de Rennes. Membre du comité de thèse.

TÉO LÉMANE (thèse en cours). *Unbiased detection of neurogenerative structural variants using k-mers matrices*, Université de Rennes. Membre du comité de thèse.

Post-doctorats

AFSHIN FAYYAZ MOVAGHAR (12 mois, 2008-2009). *Analyse statistique de la significativité des scores d'alignement séquence/structure pour la prédiction de structure 3D des protéines*, projet ANR PROTEUS, co-encadrement avec F. Rodolphe et J.-F. Gibrat.

JULIEN FAYOLLE (18 mois, 2009-2011). *Comparaison d'outils d'alignements de lectures courtes*, projet ANR CBME, co-encadrement avec J.-F. Gibrat et V. Loux.

Ingénieurs

JÉRÔME COMPAIN (18 mois, 2012-2014). *Comparaison d'outils d'alignements de lectures courtes*, projet ANR FRANCE-GENOMIQUE, co-encadrement avec V. Loux et J.-F. Gibrat.

Stages

BOSSARD, N. (1997). *Prédiction en cartographie physique par la méthode d'ancrage, en l'absence d'homogénéité sur la longueur des clones*. Mémoire de D.E.S.S. Informatique Appliquée à la Biologie, Université Versailles-Saint-Quentin (6 mois).

LEPAGE, F. (1999). *Implémentation et analyse d'une nouvelle méthode statistique de classification de protéines*. Mémoire de D.E.S.S. Informatique Appliquée à la Biologie, Université Versailles-Saint-Quentin (6 mois).

NÉDÉLEC, É. (1999). *Recherche des mots exceptionnels dans les séquences d'ADN conditionnellement à l'arrivée d'un long mot*. Mémoire de D.E.A. Modélisation Stochastique et Statistique, Université Paris-XI, Orsay (4 mois).

GUSTO, G. (2000). *Approximation par une loi de poisson composée de la loi du comptage d'un mot rare dans une chaîne de markov.* Mémoire de D.E.A. Modélisation Stochastique et Statistique, Université Paris-XI, Orsay (3 mois).

BASTIÈRE, J. (2002). *Etude de la répartition des sites Chi sur le génome de plusieurs souches de Escherichia coli*. Mémoire de Maîtrise de Biologie Cellulaire et Physiologie, Université de Versailles. (1 mois).

LAJUS, A. (2002). *Conception d'une interface JAVA pour la comparaison de l'exceptionnalité des mots dans deux séquences d'ADN*. Mémoire de 3ème année IUP Génie Biologique et Informatique, Université d'Evry. (7 mois).

WYNANT, W. (2002). *Modélisation de la corépartition de mots le long d'un génome*. Mémoire de DESS Ingénierie Mathématique - Université Paris XI. (6 mois).

DEMIZIEUX, R. (2002). *Etude de la significativité statistisque des scores obtenus à partir de la méthode de reconnaissance de repliement des protéines FROST*. Mémoire de DEA Statistique et modèles aléatoires en Economie et Finances, Université Paris VII. (4 mois).

WYNANT, W. (2003). *Recherche de mots exceptionnels dans une chaîne de Markov cachée*. Mémoire de DEA Application des Mathématiques et de l'Informatique à la Biologie - Université d'Evry (6 mois).

BOURGAIT, I. (2003). *Comparaison d'un système de réparation de l'ADN chez des souches pathogènes et non pathogènes d'E. coli*. Mémoire de DESS Ingénierie et Génomique Fonctionnelle - Université de Paris 7 (6 mois).

SIMON, G. (2003). *Analyse statistique du contexte des motifs Chi chez E. coli*. Mémoire de 1ère année de l'ENSAE - Malakoff (5 semaines).

GUÉRIN, J. (2004). *Réalisation d'une interface java pour les résultats du logiciel R'MES* Mémoire de DESS Informatique Appliquée à la Biologie - Universités Paris VI, Versailles et Evry. (5 mois).

ROQUAIN, E. (2004). *Approximation de Poisson composée du comptage d'une famille de mots dans une chaîne de Markov.* Mémoire de DEA Modélisation Stochastique et Statistique - Université d'Orsay. (5 mois).

GROSZ, S. (2006). *Indice de confiance pour les boucles spécifiques d'un génome obtenues par alignement de génomes complets.* Mémoire de Master 2ème année BioInformatique et Bio-Statistiques - Université d'Orsay. (6 mois).

THABET, H. (2007). *Analyse statistique du nombre de mots communs à deux génomes.* Mémoire de Master 2 Recherche Modélisation Aléatoire - Universtité Paris 7. (5 mois).

ARISTE, O. (2010). *Evolution du motif Chi chez Escherichia coli et Staphylococcus aureus.* Mémoire de Licence professionnelle - IUT de Perpignan. (5 mois).

BELISSEN, V., CABRAL-MORAES, J.-F., DARTEVELLE, B., DUMAZERT, J., FAUGEROUX, R. and STEPHAN, J.-Y. (2012). *Analyse de génome par une approche statistique, l'utilisation des matrices poids-position.* Projet Scientifique Collectif, 2ème année de l'Ecole Polytechnique. (8 mois).

LACOUR, H. (2012). *Comparaison de méthodes de classification de séquences d'ADN et application à des données métagénomiques.* Mémoire de 2ème année de l'Institut de Statistique de Paris. (4 mois).

PRIVAT, M. (2014). Logiques cis-régulatrices dans le développement du cerveau moyen. Master's thesis, ENS Lyon - L3. (2 mois).

AHMED ZAID, L. (2017). Analyse du lien entre puissance métabolique et entropie du génome chez les microorganismes. Master's thesis, Université de Bordeaux. (4 mois).

LAO, J. (2017). Comparaison de méthodes statistiques d'inférence de réseaux de co-occurences au sein d'écosystèmes microbiens à partir de données métagénomiques. Master's thesis, Université Paris-Diderot. (6 mois).

DULAC, G. (2020). Taxonomic classification of hyper variable regions with machine learning. Research paper. Master Data science for business. X-HEC.

23 décembre 2021.