Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods

Background
Whole genome sequencing (WGS) is increasingly used in diagnostics and surveillance of infectious diseases. A major application for WGS is to use the data for identifying outbreak clusters, and there is therefore a need for methods that can accurately and efficiently infer phylogenies from sequencing reads. In the present study we describe a new dataset that we have created for the purpose of benchmarking such WGS-based methods for epidemiological data, and also present an analysis where we use the data to compare the performance of some current methods.

Results
Our aim was to create a benchmark data set that mimics sequencing data of the sort that might be collected during an outbreak of an infectious disease. This was achieved by letting an E. coli hypermutator strain grow in the lab for 8 consecutive days, each day splitting the culture in two while also collecting samples for sequencing. The result is a data set consisting of 101 whole genome sequences with known phylogenetic relationship. Among the sequenced samples 51 correspond to internal nodes in the phylogeny because they are ancestral, while the remaining 50 correspond to leaves. We also used the newly created data set to compare three different online available methods that infer phylogenies from whole-genome sequencing reads: NDtree, CSI Phylogeny and REALPHY. One complication when comparing the output of these methods with the known phylogeny is that phylogenetic methods typically build trees where all observed sequences are placed as leaves, even though some of them are in fact ancestral. We therefore devised a method for post processing the inferred trees by collapsing short branches (thus relocating some leaves to internal nodes), and also present two new measures of tree similarity that takes into account the identity of both internal and leaf nodes.

Conclusions
Based on this analysis we find that, among the investigated methods, CSI Phylogeny had the best performance, correctly identifying 73% of all branches in the tree and 71% of all clades. We have made all data from this experiment (raw sequencing reads, consensus whole-genome sequences, as well as descriptions of the known phylogeny in a variety of formats) publicly available, with the hope that other groups may find this data useful for benchmarking and exploring the performance of epidemiological methods. All data is freely available at: https://cge.cbs.dtu.dk/services/evolution_data.php.

General information
State: Published
Organisations: Department of Bio and Health Informatics, Genomic Epidemiology, Disease Intelligence and Molecular Evolution, National Food Institute, Research Group for Genomic Epidemiology, University of Copenhagen
Authors: Ahrenfeldt, J. (Intern), Skaarup, C. (Intern), Hasman, H. (Ekstern), Pedersen, A. G. (Intern), Aarestrup, F. M. (Intern), Lund, O. (Intern)
Number of pages: 13
Publication date: 2017
Main Research Area: Technical/natural sciences

Publication information
Journal: BMC Genomics
Volume: 18
Article number: 19
ISSN (Print): 1471-2164
Ratings:
BFI (2018): BFI-level 1
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 1
Web of Science (2017): Indexed yes
BFI (2016): BFI-level 1
Scopus rating (2016): CiteScore 4.05 SJR 2.065 SNIP 1.122
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 1
Scopus rating (2015): SJR 2.287 SNIP 1.172 CiteScore 4.3
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 1
Scopus rating (2014): SJR 2.297 SNIP 1.205 CiteScore 4.18
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 1
Scopus rating (2013): SJR 2.141 SNIP 1.174 CiteScore 4.39
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
BFI (2012): BFI-level 1
Scopus rating (2012): SJR 2.181 SNIP 1.225 CiteScore 4.61
ISI indexed (2012): ISI indexed yes
Web of Science (2012): Indexed yes
BFI (2011): BFI-level 1
Scopus rating (2011): SJR 2.271 SNIP 1.197 CiteScore 4.38
ISI indexed (2011): ISI indexed yes
Web of Science (2011): Indexed yes
BFI (2010): BFI-level 1
Breadth of T cell responses after immunization with adenovirus vectors encoding ancestral antigens or polyvalent papillomavirus antigens

Oncogenic human papillomaviruses (HPVs) are in most cases eliminated by intervention of T cells. As many other pathogens, these oncogenic HPVs belong to an ancient and diverse virus family. Therefore, we found it relevant to investigate the potential and limitations of inducing a broad response - either by inducing cross-reactive T cells or by administering a polyvalent vaccine. To test these strategies, we designed 3 ancestral and 2 circulating sequences based on the two domains of the E1 and E2 proteins of papillomaviruses (PVs) that exhibit the highest degree of conservation in comparison to the other PV proteins. The PV sequences were fused to a T cell adjuvant, the murine invariant chain and encoded in a recombinant adenoviral vector which was administered to naïve outbred mice. By measuring T cell responses induced by these different vaccines and towards peptide pools representing 3 circulating strains and a putative ancestor of oncogenic HPVs, we showed that the ancestral vaccine antigen has to be approximately 90% identical to the circulating PVs before a marked drop of ~90% mean CD8+ T cell responses ensues. Interestingly, the combination of two or three type-specific PV vaccines did not induce a significant decrease of the CD8+ T cell response to the individual targeted PV types. Polyvalent HPV vaccine based on the E1 and E2 proteins seem to be capable of triggering responses towards more than one type of PV while the cross-reactivity of ancestral vaccine seems insufficient in consideration of the sequence diversity between HPV types.

General information
State: Published
Organisations: Center for Biological Sequence Analysis, Department of Bio and Health Informatics, Disease Intelligence and Molecular Evolution
Authors: Ragonnaud, E. (Ekstern), Pedersen, A. G. (Intern), Holst, P. J. (Ekstern)
Evolutionary analysis of whole-genome sequences confirms inter-farm transmission of Aleutian mink disease virus

Aleutian mink disease virus (AMDV) is a frequently encountered pathogen associated with mink farming. Previous phylogenetic analyses of AMDV have been based on shorter and more conserved parts of the genome, e.g. the partial NS1 gene. Such fragments are suitable for detection but are less useful for elucidating transmission pathways while sequencing entire viral genomes provides additional informative sites and often results in better-resolved phylogenies. We explore how whole-genome sequencing can benefit investigations of AMDV transmission by reconstructing the relationships between AMDV field samples from a Danish outbreak. We show that whole-genome phylogenies are much better resolved than those based on the partial NS1 gene sequences extracted from the same alignment. Well-resolved phylogenies contain more information about the underlying transmission trees and are useful for understanding the spread of a pathogen. In the main case investigated here, the transmission path suggested by the tree structure was supported by epidemiological data. The use of molecular clock models further improved tree resolution and provided time estimates for the viral ancestors consistent with the proposed direction of spread. It was however impossible to infer transmission pathways from the partial NS1 gene tree, since all samples from the case farms branched out from a single internal node.

A sliding window analysis showed that there were no shorter genomic regions providing the same phylogenetic resolution as the entire genome. Altogether, these results suggest that phylogenetic analyses based on whole-genome sequencing taking into account sampling dates and epidemiological data is a promising set of tools for clarifying AMDV transmission.
Scopus rating (2011): SJR 1.675 SNIP 1.149 CiteScore 3.6
ISI indexed (2011): ISI indexed yes
Web of Science (2011): Indexed yes
BFI (2010): BFI-level 1
Scopus rating (2010): SJR 1.657 SNIP 1.058
Web of Science (2010): Indexed yes
BFI (2009): BFI-level 1
Scopus rating (2009): SJR 1.644 SNIP 1.13
Web of Science (2009): Indexed yes
BFI (2008): BFI-level 1
Scopus rating (2008): SJR 1.636 SNIP 1.068
Web of Science (2008): Indexed yes
Scopus rating (2007): SJR 1.581 SNIP 1.135
Web of Science (2007): Indexed yes
Scopus rating (2006): SJR 1.688 SNIP 1.127
Web of Science (2006): Indexed yes
Scopus rating (2005): SJR 1.626 SNIP 1.138
Web of Science (2005): Indexed yes
Scopus rating (2004): SJR 1.524 SNIP 1.201
Web of Science (2004): Indexed yes
Scopus rating (2003): SJR 1.568 SNIP 1.146
Web of Science (2003): Indexed yes
Scopus rating (2002): SJR 1.428 SNIP 1.085
Web of Science (2002): Indexed yes
Scopus rating (2001): SJR 1.486 SNIP 1.115
Web of Science (2001): Indexed yes
Scopus rating (2000): SJR 1.407 SNIP 1.065
Web of Science (2000): Indexed yes
Scopus rating (1999): SJR 1.374 SNIP 1.079
Original language: English
Aleutian mink disease virus (AMDV), Whole-genome sequencing, Next-generation sequencing (NGS), Phylogeny, Viral outbreak investigation
DOIs:
10.1099/jgv.0.000777
Source: FindIt
Source-ID: 2371489476
Publication: Research - peer-review › Journal article – Annual report year: 2017

Forskellige virusstammer var årsag til udbred af plasmacytose i danske mink (Neovison vison) i 2015

General information
State: Published
Organisations: National Veterinary Institute, Virology, Department of Bio and Health Informatics, Disease Intelligence and Molecular Evolution, Diagnostic & Development, Department of Biotechnology and Biomedicine, Kopenhagen Fur
Pages: 163-167
Publication date: 2017

Host publication information
Title of host publication: Faglig årsberetning 2016 : Kopenhagen Fur
Place of publication: Aarhus N
Publisher: Kopenhagen Fur
Main Research Area: Technical/natural sciences
Electronic versions:
DTUS.pdf
Publication: Research - peer-review › Book chapter – Annual report year: 2017
Molecular diagnostics of aleutian mink disease virus: applied use of next generation sequencing and phylogenetics

Aleutian Mink Disease virus (AMDV) is a parvovirus causing Aleutian Mink Disease (AMD), often referred to as plasmacytosis. It is a systemic infection affecting mink of all ages, and is globally the most important pathogen impacting mink farming. In Denmark AMDV has since 1999 been monitored by a national control program, which is based on serological screening of all animals and encourages infected farms to stamp out. Historically there has been no consensus about which genomic region of the virus to analyse e.g. in relation to surveillance, and most previous studies in this regard, have been based either on partial or entire genes, or on pure epidemiological data. Thus, when initiating this project, little was known about AMDV's total genomic diversity and how the virus was spread between farms.

Recent advances in the field of molecular diagnostics have made high throughput tools such as next generation sequencing cheaper and more easily available. Whole genome sequencing and advanced phylogenetic analyses have successfully been applied to describe the molecular evolution and transmission patterns for viruses such as Foot and Mouth Disease Virus (FMDV), Ebola, and avian influenza virus, however not previously for AMDV. The overall aim with this thesis was to investigate if next generation sequencing and phylogenetic analyses of full length isolates could improve our understanding of the total genomic diversity and evolution of AMDV. Additionally, we wanted to evaluate if this knowledge could contribute to the elucidation of AMDV transmission between farms and improve molecular diagnostics. During the first phase of this project a method for performing whole genome sequencing of AMDV was developed. This protocol enabled the sequencing of a large number of in vivo infectious AMDV isolates and provided the necessary dataset to act as foundation for the remaining analyses in the thesis. The first original paper (Manuscript 1) describes this protocol.

Manuscript 2 is a proof-of-concept study which demonstrated the advantage of using the whole genome sequence approach, compared to the in Denmark traditionally used partial NS1 gene sequencing, for the elucidation of transmission pathways between farms. The study has been performed on samples from a small local AMDV outbreak, and clearly illustrated that the phylogenies based on partial NS1 gene sequencing were uninformative and could not be used for determining transmission pathways, even in the light of supporting epidemiological data. The whole-genome approach on the other hand, confirmed the epidemiological hypothesis about the direction of spread.

In Manuscript 3, the methodologies from Manuscript 1 and 2 were applied to generate the to-date most comprehensive phylogenetic and genetic analysis of full-length AMDV isolates, composed of more than 200 field strains. The study shed light on the diversity and evolutionary behaviour of two distinct AMDV strains, in addition to providing the first robust evolutionary rate-estimates. Altogether, the work presented in this thesis provides a contribution to the molecular diagnostics of AMDV, enables us better to understand the virus' evolutionary behaviour in the context of mink farming, and is anticipated to be of value for more accurately tracing back in time the emergence of future outbreaks.
Niche differentiation and evolution of comammox Nitrospira through a comparative genomics analysis

Nitrification, the biological oxidation of ammonium to nitrate, is a fundamental process in the nitrogen cycle and plays an important role in natural and engineered systems. Throughout the last century, nitrification was assumed to be a two-step process executed by two different functional groups, ammonia oxidizing prokaryotes (AOP) and nitrite oxidizing bacteria (NOB). Recently, several articles have shown the capability of a single microorganism, belonging to the genus Nitrospira, to carry out the complete oxidation of ammonia to nitrate (comammox). Nitrospira spp. are widespread in both natural and engineered ecosystems associated with nitrogen cycling and different species are frequently observed to coexist in the same environment. Besides recent discoveries pointing towards versatile metabolism in some Nitrospira species, little is known about the functional potential of the two comammox Nitrospira clades, and the factors involved in niche-partitioning between comammox and canonical Nitrospira.

A comparative genomics analysis was conducted with five genomes recovered from a groundwater-fed rapid sand filter (including both comammox clades and a nitrite-oxidizing Nitrospira population genome) and high quality published Nitrospira genomes, to reveal distinct genomic features within Nitrospira. In addition, we investigated the evolution of the ammonia oxidation pathway in comammox Nitrospira. This analysis revealed distinct genetic capabilities of the different comammox clades and canonical Nitrospira which can help to explain the coexistence and niche partitioning of Nitrospira spp. These divergences range from the nitrogen source utilization capacity to the ability for electron donor versatility, and other characteristics such as stress response. With respect to the evolutionary history of comammox Nitrospira, our analysis indicates transfer events between betaproteobacterial ammonia oxidizers. In addition, transfer events between comammox clade A and clade B were also detected for genes belonging to the ammonium oxidation pathway. Together, these results expand the actual knowledge of the ecology and evolution of the recently discovered comammox Nitrospira.

Therapeutic Vaccine Against Primate Papillomavirus Infections of the Cervix

Currently available prophylactic vaccines have no therapeutic efficacy for preexisting human papillomavirus (HPVs) infections, do not target all oncogenic HPVs and are insufficient to eliminate the burden of HPV induced cancer. We aim to develop an alternative HPV vaccine which is broadly effective and capable of clearing preexisting infection. In an initial attempt to develop a broadly reactive therapeutic vaccine, we designed a putative papillomavirus (PV) ancestor antigen (circulating sequence derived antigenic sequences E1E2-CDSE1E2) based on the conserved E1 and E2 protein sequences from existing oncogenic HPV strains. This antigen was found to be as related to circulating oncogenic Macaca fascicularis papillomaviruses (MfPVs) as to oncogenic HPVs. The CDSE1E2 antigen was fused to a T-cell adjuvant and encoded in chimpanzee 3 and 63 adenoviral vectors. We first showed that the combination of these 2 vaccines induced long-lasting potent CDSE1E2 specific T cell responses in outbred mice. This prime-boost regimen was then tested in female macaques naturally infected with MfPVs. All immunized animals (16/16) responded to the vaccine antigen but with reduced cross-reactivity against existing PVs. Preexisting MFPV infections did not prime vaccine inducible immune responses. Importantly, immunized oncogenic MFPV type 3 (MFPV3) infected animals that responded toward MFPV3 were able to diminish cervical MFPV3 DNA content. Although insufficient breadth was achieved, our results suggest that a relevant level of E1E2 specific T cell immunity is achievable and might be sufficient for the elimination of PV infection. Importantly, naturally infected macaques, offer a relevant model for testing vaccines aimed at eliminating mucosal PV infections.
Aleutian Mink Disease Virus (AMDV) is a frequently encountered pathogen associated with commercial mink breeding. AMDV infection leads to increased mortality and compromised animal health and welfare. Currently little is known about the molecular evolution of the virus, and the few existing studies have focused on limited regions of the viral genome. This
paper describes a robust, reliable, and fast protocol for amplification of the full AMDV genome using long-range PCR. The method was used to generate next generation sequencing data for the non-virulent cell-culture adapted AMDV-G strain as well as for the virulent AMDV-Utah strain. Comparisons at nucleotide- and amino acid level showed that, in agreement with existing literature, the highest variability between the two virus strains was found in the left open reading frame, which encodes the non-structural (NS1–3) genes. This paper also reports a number of differences that potentially can be linked to virulence and host range. To the authors’ knowledge, this is the first study to apply next generation sequencing on the entire AMDV genome. The results from the study will facilitate the development of new diagnostic tools and can form the basis for more detailed molecular epidemiological analyses of the virus.

General information
State: Published
Organisations: Department of Systems Biology, Molecular Evolution, National Veterinary Institute, Section for Virology, Center for Biological Sequence Analysis, Kopenhagen Fur
Number of pages: 9
Pages: 43-51
Publication date: 2016
Main Research Area: Technical/natural sciences

Publication information
Journal: Journal of Virological Methods
Volume: 234
ISSN (Print): 0166-0934
Ratings:
BFI (2018): BFI-level 1
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 1
Web of Science (2017): Indexed Yes
BFI (2016): BFI-level 1
Scopus rating (2016): SJR 0.87 SNIP 0.736 CiteScore 1.78
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 1
Scopus rating (2015): SJR 0.868 SNIP 0.799 CiteScore 1.68
BFI (2014): BFI-level 1
Scopus rating (2014): SJR 0.893 SNIP 0.952 CiteScore 1.87
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 1
Scopus rating (2013): SJR 0.861 SNIP 0.91 CiteScore 1.99
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
BFI (2012): BFI-level 1
Scopus rating (2012): SJR 0.869 SNIP 0.935 CiteScore 2.08
ISI indexed (2012): ISI indexed yes
Web of Science (2012): Indexed yes
BFI (2011): BFI-level 1
Scopus rating (2011): SJR 0.907 SNIP 0.994 CiteScore 2.23
ISI indexed (2011): ISI indexed yes
Web of Science (2011): Indexed yes
BFI (2010): BFI-level 1
Scopus rating (2010): SJR 0.892 SNIP 0.998
Web of Science (2010): Indexed yes
BFI (2009): BFI-level 1
Scopus rating (2009): SJR 0.964 SNIP 1.061
Web of Science (2009): Indexed yes
BFI (2008): BFI-level 1
Scopus rating (2008): SJR 0.918 SNIP 1.082
Web of Science (2008): Indexed yes
Fish Oil-Derived Fatty Acids in Pregnancy and Wheeze and Asthma in Offspring

Background: Reduced intake of n-3 long-chain polyunsaturated fatty acids (LCPUFAs) may be a contributing factor to the increasing prevalence of wheezing disorders. We assessed the effect of supplementation with n-3 LCPUFAs in pregnant women on the risk of persistent wheeze and asthma in their offspring. Methods: We randomly assigned 736 pregnant women at 24 weeks of gestation to receive 2.4 g of n-3 LCPUFA (fish oil) or placebo (olive oil) per day. Their children formed the Copenhagen Prospective Studies on Asthma in Childhood2010 (COPSAC2010) cohort and were followed prospectively with extensive clinical phenotyping. Neither the investigators nor the participants were aware of group assignments during follow-up for the first 3 years of the children's lives, after which there was a 2-year follow-up period during which only the investigators were unaware of group assignments. The primary end point was persistent wheeze or asthma, and the secondary end points included lower respiratory tract infections, asthma exacerbations, eczema, and allergic sensitization.

Results: A total of 695 children were included in the trial, and 95.5% completed the 3-year, double-blind follow-up period. The risk of persistent wheeze or asthma in the treatment group was 16.9%, versus 23.7% in the control group (hazard ratio, 0.69; 95% confidence interval [CI], 0.49 to 0.97; P=0.035), corresponding to a relative reduction of 30.7%. Prespecified subgroup analyses suggested that the effect was strongest in the children of women whose blood levels of eicosapentaenoic acid and docosahexaenoic acid were in the lowest third of the trial population at randomization: 17.5% versus 34.1% (hazard ratio, 0.46; 95% CI, 0.25 to 0.83; P=0.011). Analyses of secondary end points showed that supplementation with n-3 LCPUFA was associated with a reduced risk of infections of the lower respiratory tract (31.7% vs. 39.1%; hazard ratio, 0.75; 95% CI, 0.58 to 0.98; P=0.033), but there was no statistically significant association between supplementation and asthma exacerbations, eczema, or allergic sensitization.

Conclusions: Supplementation with n-3 LCPUFA in the third trimester of pregnancy reduced the absolute risk of persistent wheeze or asthma and infections of the lower respiratory tract in offspring by approximately 7 percentage points, or one third. (Funded by the Lundbeck Foundation and others; ClinicalTrials.gov number, NCT00798226.)

General information
State: Published
Organisations: Department of Systems Biology, Center for Biological Sequence Analysis, University of Copenhagen, Statens Serum Institut, University of Waterloo
Outbreaks of Aleutian mink disease in farmed mink (Neovison vison) in Denmark: molecular characterization by partial NS1 gene sequencing

General information
State: Published
Organisations: National Veterinary Institute, Section for Virology, Department of Systems Biology, Center for Biological Sequence Analysis, Molecular Evolution, Department of Bio and Health Informatics, Kopenhagen Fur
Pages: 85-87
Publication date: 2016

Host publication information
Title of host publication: Proceedings of the XIth International Scientific Congress in Fur Animal Production
Place of publication: Helsinki, Finland
Publisher: Libris
Editors: Mäki-Tanila, A., Valaja, J., Mononen, J., Sironen, T., Vapalahti, O.
Series: Scientifur
Volume: 40
Number: 3/4
ISSN: 2445-6292
Main Research Area: Technical/natural sciences
Electronic versions:
IFASA2016_Vol.40_1_.pdf
Links:
Publication: Research - peer-review › Conference abstract in proceedings – Annual report year: 2016

Using expected sequence features to improve basecalling accuracy of amplicon pyrosequencing data
Amplicon pyrosequencing targets a known genetic region and thus inherently produces reads highly anticipated to have certain features, such as conserved nucleotide sequence, and in the case of protein coding DNA, an open reading frame. Pyrosequencing errors, consisting mainly of nucleotide insertions and deletions, are on the other hand likely to disrupt open reading frames. Such an inverse relationship between errors and expectation based on prior knowledge can be used advantageously to guide the process known as basecalling, i.e. the inference of nucleotide sequence from raw sequencing data. The new basecalling method described here, named Multipass, implements a probabilistic framework for working with the raw flowgrams obtained by pyrosequencing. For each sequence variant Multipass calculates the likelihood and nucleotide sequence of several most likely sequences given the flowgram data. This probabilistic approach enables integration of basecalling into a larger model where other parameters can be incorporated, such as the likelihood for observing a full-length open reading frame at the targeted region. We apply the method to 454 amplicon pyrosequencing data obtained from a malaria virulence gene family, where Multipass generates 20 % more error-free sequences than current state of the art methods, and provides sequence characteristics that allow generation of a set of high confidence error-free sequences. This novel method can be used to increase accuracy of existing and future amplicon sequencing data, particularly where extensive prior knowledge is available about the obtained sequences, for example in analysis of the immunoglobulin VDJ region where Multipass can be combined with a model for the known recombining germline genes. Multipass is available for Roche 454 data at http://www.cbs.dtu.dk/services/MultiPass-1.0 , and the concept can potentially be implemented for other sequencing technologies as well.

General information
State: Published
Organisations: Department of Systems Biology, Center for Biological Sequence Analysis, Metagenomics, Molecular Evolution, New York University
Authors: Rask, T. S. (Intern), Petersen, B. (Intern), Chen, D. S. (Ekstern), Day, K. P. (Ekstern), Pedersen, A. G. (Intern)
Number of pages: 10
Publication date: 2016
Main Research Area: Technical/natural sciences

Publication information
Journal: BMC Bioinformatics
Volume: 17
Issue number: 1
Article number: 176
Creation of Functional Viruses from Non-Functional cDNA Clones Obtained from an RNA Virus Population by the Use of Ancestral Reconstruction

RNA viruses have the highest known mutation rates. Consequently it is likely that a high proportion of individual RNA virus genomes, isolated from an infected host, will contain lethal mutations and be non-functional. This is problematic if the aim is to clone and investigate high-fitness, functional cDNAs and may also pose problems for sequence-based analysis of viral evolution. To address these challenges we have performed a study of the evolution of classical swine fever virus (CSFV) using deep sequencing and analysis of 84 full-length cDNA clones, each representing individual genomes from a moderately virulent isolate. In addition to here being used as a model for RNA viruses generally, CSFV has high socioeconomic importance and remains a threat to animal welfare and pig production. We find that the majority of the investigated genomes are non-functional and only 12% produced infectious RNA transcripts. Full length sequencing of cDNA clones and deep sequencing of the parental population identified substitutions important for the observed phenotypes. The investigated cDNA clones were furthermore used as the basis for inferring the sequence of functional viruses. Since each unique clone must necessarily be the descendant of a functional ancestor, we hypothesized that it should be possible to produce functional clones by reconstructing ancestral sequences. To test this we used phylogenetic methods to infer two ancestral sequences, which were then reconstructed as cDNA clones. Viruses rescued from the reconstructed cDNAs were tested in cell culture and pigs. Both reconstructed ancestral genomes proved functional, and displayed distinct phenotypes in vitro and in vivo. We suggest that reconstruction of ancestral viruses is a useful tool for experimental and computational investigations of virulence and viral evolution. Importantly, ancestral reconstruction can be done even on the basis of a set of sequences that all correspond to non-functional variants.
Early Divergent Strains of *Yersinia pestis* in Eurasia 5,000 Years Ago

The bacteria *Yersinia pestis* is the etiological agent of plague and has caused human pandemics with millions of deaths in historic times. How and when it originated remains contentious. Here, we report the oldest direct evidence of *Yersinia pestis* identified by ancient DNA in human teeth from Asia and Europe dating from 2,800 to 5,000 years ago. By sequencing the genomes, we find that these ancient plague strains are basal to all known *Yersinia pestis*. We find the origins of the *Yersinia pestis* lineage to be at least two times older than previous estimates. We also identify a temporal sequence of genetic changes that lead to increased virulence and the emergence of the bubonic plague. Our results show that plague infection was endemic in the human populations of Eurasia at least 3,000 years before any historical recordings of pandemics.

**General information**

State: Published
Organisations: Department of Systems Biology, Center for Biological Sequence Analysis, Metagenomics, Molecular Evolution, Eucaryotic Molecular Cell Biology, Integrative Systems Biology, University of Copenhagen, National Academy of Sciences of the Republic of Armenia, South Ural State University, Orenburg Museum of Fine Arts, Yerevan State University, University of Tartu, University of Wroclaw, Peter the Great Museum of Anthropology and Ethnography, University of Cambridge, University of California, University of Gothenburg
Number of pages: 12
Pages: 571-582
DNA secondary structures are associated with recombination in major Plasmodium falciparum variable surface antigen gene families

Many bacterial, viral and parasitic pathogens undergo antigenic variation to counter host immune defense mechanisms. In Plasmodium falciparum, the most lethal of human malaria parasites, switching of var gene expression results in alternating expression of the adhesion proteins of the Plasmodium falciparum-erythrocyte membrane protein 1 class on the infected erythrocyte surface. Recombination clearly generates var diversity, but the nature and control of the genetic exchanges involved remain unclear. By experimental and bioinformatic identification of recombination events and genome-wide recombination hotspots in var genes, we show that during the parasite’s sexual stages, ectopic recombination between isogenous var paralogs occurs near low folding free energy DNA 50-mers and that these sequences are heavily concentrated at the boundaries of regions encoding individual Plasmodium falciparum-erythrocyte membrane protein 1 structural domains. The recombinogenic potential of these 50-mers is not parasite-specific because these sequences also induce recombination when transferred to the yeast Saccharomyces cerevisiae. Genetic cross data suggest that DNA secondary structures (DSS) act as inducers of recombination during DNA replication in P. falciparum sexual stages, and that these DSS-regulated genetic exchanges generate functional and diverse P. falciparum adhesion antigens. DSS-induced recombination may represent a common mechanism for optimizing the evolvability of virulence gene families in pathogens.

General information
State: Published
Organisations: Department of Systems Biology, Center for Biological Sequence Analysis, University of Edinburgh, Weill Cornell Medical College, University of Copenhagen
Pages: 2270-2281
Publication date: 2014
Main Research Area: Technical/natural sciences

Publication information
Journal: Nucleic Acids Research
Volume: 42
Issue number: 4
ISSN (Print): 0305-1048
Ratings:
BFI (2018): BFI-level 2
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 2
Web of Science (2017): Indexed yes
BFI (2016): BFI-level 2
Scopus rating (2016): CiteScore 9.28 SJR 7.397 SNIP 2.657
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 2
Scopus rating (2015): SJR 7.239 SNIP 2.639 CiteScore 9.48
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 2
Scopus rating (2014): SJR 6.576 SNIP 2.568 CiteScore 8.74
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 2
Scopus rating (2013): SJR 6.582 SNIP 2.266 CiteScore 8.46
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
BFI (2012): BFI-level 2
Scopus rating (2012): SJR 6.13 SNIP 2.392 CiteScore 8.62
ISI indexed (2012): ISI indexed yes
Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes

Most current approaches for analyzing metagenomic data rely on comparisons to reference genomes, but the microbial diversity of many environments extends far beyond what is covered by reference databases. De novo segregation of complex metagenomic data into specific biological entities, such as particular bacterial strains or viruses, remains a largely unsolved problem. Here we present a method, based on binning co-abundant genes across a series of metagenomic samples, that enables comprehensive discovery of new microbial organisms, viruses and co-inherited genetic entities and aids assembly of microbial genomes without the need for reference sequences. We demonstrate the method on data from 396 human gut microbiome samples and identify 7,381 co-abundance gene groups (CAGs), including 741 metagenomic species (MGS). We use these to assemble 238 high-quality microbial genomes and identify affiliations between MGS and hundreds of viruses or genetic entities. Our method provides the means for comprehensive profiling of the diversity within complex metagenomic samples.

General information
State: Published
Organisations: Department of Systems Biology, Center for Biological Sequence Analysis, Novo Nordisk Foundation Center for Biosustainability, Center for Biological sequence analysis, Comparative Microbial Genomics, National Food Institute, Division of Epidemiology and Microbial Genomics, INRA Institut National de la Recherche Agronomique, South China University of Technology, European Molecular Biology Laboratory, Centre National de la Recherche Scientifique, University of Southern Denmark, University Hospital Vall d’Hebron, University of Copenhagen, Vrije Universiteit Brussel,
Rescue of the highly virulent classical swine fever virus strain "Koslov" from cloned cDNA and first insights into genome variations relevant for virulence

Classical swine fever virus (CSFV) strain "Koslov" is highly virulent with a mortality rate of up to 100% in pigs. In this study, we modified non-functional cDNAs generated from the blood of Koslov virus infected pigs by site-directed mutagenesis, removing non-synonymous mutations step-by-step, thereby producing genomes encoding the consensus amino acid sequence. Viruses rescued from the construct corresponding to the inferred parental form were highly virulent, when tested in pigs, with infected animals displaying pronounced clinical symptoms leading to high mortality. The reconstruction therefore gave rise to a functional cDNA corresponding to the highly virulent Koslov strain of CSFV. It could be demonstrated that two single amino acid changes (S763L and P968H) in the surface structural protein E2 resulted in attenuation in the porcine infection system while another single amino acid change within the nonstructural protein NS3 (D2183G) reduced virus growth within cells in vitro.

General information
State: Published
Organisations: Molecular Evolution, National Veterinary Institute, Section for Virology, Systems Biotechnology, Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, Friedrich Loeffler Institute
Authors: Fahnøe, U. (Intern), Pedersen, A. G. (Intern), Risager, P. C. (Intern), Nielsen, J. (Ekstern), Belsham, G. (Intern), Höper, D. (Ekstern), Beer, M. (Ekstern), Rasmussen, T. B. (Intern)
Number of pages: 9
Pages: 379-387
Publication date: 2014
Main Research Area: Technical/natural sciences

Publication information
Journal: Virology
Volume: 468-470
ISSN (Print): 0042-6822
Ratings:
BFI (2018): BFI-level 1
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 1
Web of Science (2017): Indexed Yes
BFI (2016): BFI-level 1
Scopus rating (2016): SJR 1.883 SNIP 0.96 CiteScore 3.47
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 1
Scopus rating (2015): SJR 1.789 SNIP 0.898 CiteScore 3.2
BFI (2014): BFI-level 1
Scopus rating (2014): SJR 1.695 SNIP 0.919 CiteScore 3.14
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 1
Targeting the genetic complexity within adapting RNA virus populations

General information
State: Published
Organisations: Molecular Evolution, National Veterinary Institute, Section for Virology, Department of Systems Biology, Center for Biological Sequence Analysis
Authors: Fahnøe, U. (Intern), Pedersen, A. G. (Intern), Rasmussen, T. B. (Intern)
Number of pages: 160
Publication date: 2014

Publication information
Publisher: Technical University of Denmark (DTU)
Original language: English
Main Research Area: Technical/natural sciences
Electronic versions:
PhD_Thesis_Ulrik_Fahn_e.pdf
Source: PublicationPreSubmission
Source-ID: 108130125
Publication: Research › Ph.D. thesis – Annual report year: 2015
Comparison of two Next Generation sequencing platforms for full genome sequencing of Classical Swine Fever Virus

Next Generation Sequencing (NGS) is becoming more adopted into viral research and will be the preferred technology in the years to come. We have recently sequenced several strains of Classical Swine Fever Virus (CSFV) by NGS on both Genome Sequencer FLX (GS FLX) and Iontorrent PGM platforms. In this study, we analyzed NGS data of virus rescued from a CSFV C-strain vaccine strain cDNA clone. The virus analyzed was obtained from a 4th and a 12th passage of rescued virus in SFT cell culture, which had shown a difference in growth kinetics between the passages, and NGS analysis was chosen in order to look for molecular differences. Identical RT-PCR products were run on both GS FLX and an Iontorrent PGM platform for comparison. The NGS data was compared by quality and the percentage mapped reads. Results showed good quality of reads for both platforms and a close to 100% of the reads mapped to the consensus sequence. Additionally, we got an average sequence depth for the genome of 4000 for the Iontorrent PGM and 400 for the FLX platform making the mapping suitable for single nucleotide variant (SNV) detection. The analysis revealed a single non-silent SNV A10665G leading to the amino acid change D3431G in the RNAdependent RNA polymerase NS5B. This SNV was present at 100% frequency in the 12th passage and only at 55% in the 4th passage, which could explain the difference in growth kinetics between the passages.

Reconstructing the highly virulent Classical Swine Fever Virus strain Koslov

Classical swine fever virus (CSFV) may be highly virulent in pigs with a mortality rate close to 100%. The CSFV “Koslov strain” is known to be one of the most virulent CSFV, but so far a functional cloned cDNA of this strain has not been described. We suggest that this may be due to the error-prone nature of the RNA-dependent RNA polymerase resulting in the majority of circulating forms being non-functional. However, since any infectious virus particle should necessarily be the offspring of a functional virus, we hypothesized that it should be possible to synthesize a highly virulent form by reconstructing ancestral sequences. To test this hypothesis, we inferred sequences that correspond to ancestral nodes in a phylogenetic tree built from full-length nucleotide sequences of non-functional Koslov cDNAs and then proceeded to test the reconstructions. Specifically, we altered a non-functional cDNA by site directed mutagenesis, removing non-synonymous mutations step by step. In vitro testing of modified constructs did indeed lead to fully functional viruses with similar growth kinetics as the wild-type strain. Moreover, viruses rescued from the construct had the ancestral amino acid sequence and, when tested in pigs, were at least as virulent as the Koslov strain. The ancestral reconstruction therefore proved to give rise to a functional cDNA of the highly virulent Koslov strain. In vivo studies confirmed our methods and enabled us to identify nucleotide positions within the viral genome important for virulence.
Bayesian prediction of bacterial growth temperature range based on genome sequences

Background: The preferred habitat of a given bacterium can provide a hint of which types of enzymes of potential industrial interest it might produce. These might include enzymes that are stable and active at very high or very low temperatures. Being able to accurately predict this based on a genomic sequence, would thus allow for an efficient and targeted search for production organisms, reducing the need for culturing experiments. Results: This study found a total of 40 protein families useful for distinction between three thermophilicity classes (thermophiles, mesophiles and psychrophiles). The predictive performance of these protein families were compared to those of 87 basic sequence features (relative use of amino acids and codons, genomic and 16S rDNA AT content and genome size). When using naive Bayesian inference, it was possible to correctly predict the optimal temperature range with a Matthews correlation coefficient of up to 0.68. The best predictive performance was always achieved by including protein families as well as structural features, compared to either of these alone. A dedicated computer program was created to perform these predictions. Conclusions: This study shows that protein families associated with specific thermophilicity classes can provide effective input data for thermophilicity prediction, and that the naive Bayesian approach is effective for such a task. The program created for this study is able to efficiently distinguish between thermophilic, mesophilic and psychrophilic adapted bacterial genomes.
Bioinformatics approaches to malaria

Malaria is a life threatening disease found in tropical and subtropical regions of the world. Each year it kills 781 000 individuals; most of them are children under the age of five in sub-Saharan Africa. The most severe form of malaria in humans is caused by the parasite Plasmodium falciparum, which is the subject of the first part of this thesis.

The PfEMP1 protein which is encoded by the highly variable var gene family is important in the pathogenesis and immune evasion of malaria parasites. We analyzed and classified these genes based on the upstream sequence in seven Plasmodium falciparum clones. We show that the amount of nucleotide diversity is just as big within each clone as it is between the clones.

DNA methylation is an important epigenetic mark in many eukaryotic species. We are studying DNA methylation in the malaria parasite Plasmodium falciparum. The work is still in progress and will be introduced here.

One of the biggest concerns regarding the treatment of malaria is the continued development of resistance to existing drugs. Therefore, new drugs will be needed in the future. The ApiAP2 proteins are a recently discovered family of putative transcription factors. As they might perform important regulatory functions in the parasite, they could be useful as drug targets. Here, we study one of these proteins and describe our work on identifying small compounds that can interfere with its DNA binding abilities.

Specific binding of short peptides by proteins of the major histocompatibility complex (MHC) is an important event in the activation of immune responses to various pathogens. The set of peptides that can bind a specific MHC molecule can be characterized by a binding motif. In the second part of this thesis, we developed an algorithm that can distinguish several binding motifs within a mixture of peptides from different motifs.
Experience with the use of online lectures, video Modules, and wiki-websites in engineering education

We here present our experience with three computer-based teaching methodologies that we have used for a number of years in engineering education: Online lectures, video modules, and wiki websites. The aim is to provide the reader with concrete tools that can be used directly in teaching situations, and to inspire further use of information technologies.

General information
State: Published
Organisations: Center for Biological Sequence Analysis, Department of Systems Biology
Authors: Pedersen, A. G. (Intern), Wernersson, R. (Intern)
Publication date: 2011

Host publication information
Title of host publication: 7th International CDIO Conference 2011
Publisher: Technical University of Denmark (DTU)
Main Research Area: Technical/natural sciences
Conference: 7th International CDIO Conference, Copenhagen, Denmark, 20/06/2011 - 20/06/2011
Links:
http://www.cdio2011.dtu.dk/
Source: orbit
Source-ID: 279437
Publication: Research - peer-review › Conference abstract in proceedings – Annual report year: 2011

Plasmodium falciparum Erythrocyte Membrane Protein 1 Diversity in Seven Genomes – Divide and Conquer

The var gene encoded hyper-variable Plasmodium falciparum erythrocyte membrane protein 1 (PfEMP1) family mediates cytoadhesion of infected erythrocytes to human endothelium. Antibodies blocking cytoadhesion are important mediators of malaria immunity acquired by endemic populations. The development of a PfEMP1 based vaccine mimicking natural acquired immunity depends on a thorough understanding of the evolved PfEMP1 diversity, balancing antigenic variation against conserved receptor binding affinities. This study redefines and reclassifies the domains of PfEMP1 from seven genomes. Analysis of domains in 399 different PfEMP1 sequences allowed identification of several novel domain classes, and a high degree of PfEMP1 domain compositional order, including conserved domain cassettes not always associated with the established group A–E division of PfEMP1. A novel iterative homology block (HB) detection method was applied, allowing identification of 628 conserved minimal PfEMP1 building blocks, describing on average 83% of a PfEMP1 sequence. Using the HBs, similarities between domain classes were determined, and Duffy binding-like (DBL) domain subclasses were found in many cases to be hybrids of major domain classes. Related to this, a recombination hotspot was uncovered between DBL subdomains S2 and S3. The VarDom server is introduced, from which information on domain classes and homology blocks can be retrieved, and new sequences can be classified. Several conserved sequence elements were found, including: (1) residues conserved in all DBL domains predicted to interact and hold together the three DBL subdomains, (2) potential integrin binding sites in DBLα domains, (3) an acylation motif conserved in group A var genes suggesting N-terminal N-myristoylation, (4) PfEMP1 inter-domain regions proposed to be elastic disordered structures, and (5) several conserved predicted phosphorylation sites. Ideally, this comprehensive categorization of PfEMP1 will provide a platform for future studies on var/PfEMP1 expression and function.

General information
State: Published
Organisations: Center for Biological Sequence Analysis, Department of Systems Biology, University of Copenhagen
Authors: Rask, T. S. (Intern), Hansen, D. A. (Intern), Theander, T. G. (Ekstern), Pedersen, A. G. (Intern), Lavstsen, T. (Ekstern)
Publication date: 2010
Main Research Area: Technical/natural sciences

Publication information
Journal: P L o S Computational Biology (Online)
Volume: 6
Issue number: 9
ISSN (Print): 1553-7358
Ratings:
BFI (2018): BFI-level 1
Evolution of the leukotoxin promoter in genus Mannheimia

Background: The Mannheimia species encompass a wide variety of bacterial lifestyles, including opportunistic pathogens and commensals of the ruminant respiratory tract, commensals of the ovine rumen, and pathogens of the ruminant integument. Here we present a scenario for the evolution of the leukotoxin promoter among representatives of the five species within genus Mannheimia. We also consider how the evolution of the leukotoxin operon fits with the evolution and maintenance of virulence. Results: The alignment of the intergenic regions upstream of the leukotoxin genes showed significant sequence and positional conservation over a 225-bp stretch immediately proximal to the transcriptional start site of the lktC gene among all Mannheimia strains. However, in the course of the Mannheimia genome evolution, the acquisition of individual noncoding regions upstream of the conserved promoter region has occurred. The rate of evolution estimated branch by branch suggests that the conserved promoter may be affected to different extents by the types of natural selection that potentially operate in regulatory regions. Tandem repeats upstream of the core promoter were confined to M. haemolytica with a strong association between the sequence of the repeat units, the number of repeat units per promoter, and the phylogenetic history of this species. Conclusion: The mode of evolution of the intergenic regions upstream of the leukotoxin genes appears to be highly dependent on the lifestyle of the bacterium. Transition from
Avirulence to virulence has occurred at least once in M. haemolytica with some evolutionary success of bovine serotype A1/A6 strains. Our analysis suggests that changes in cis-regulatory systems have contributed to the derived virulence phenotype by allowing phase-variable expression of the leukotoxin protein. We propose models for how phase shifting and the associated virulence could facilitate transmission to the nasopharynx of new hosts.

**General information**

State: Published
Organisations: Center for Biological Sequence Analysis, Department of Systems Biology
Authors: Larsen, J. (Ekstern), Pedersen, A. G. (Intern), Davies, R. L. (Ekstern), Kuhnert, P. (Ekstern), Frey, J. (Ekstern), Christensen, H. (Ekstern), Bisgaard, M. (Ekstern), Olsen, J. E. (Ekstern)
Pages: 121
Publication date: 2009
Main Research Area: Technical/natural sciences

**Publication information**

Journal: BMC Evolutionary Biology
Volume: 9
ISSN (Print): 1471-2148
Ratings:
BFI (2018): BFI-level 1
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 1
Web of Science (2017): Indexed Yes
BFI (2016): BFI-level 1
Scopus rating (2016): SJR 1.806 SNIP 1.266 CiteScore 3.12
BFI (2015): BFI-level 1
Scopus rating (2015): SJR 2.057 SNIP 1.174 CiteScore 3.37
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 1
Scopus rating (2014): SJR 2.2 SNIP 1.268 CiteScore 3.42
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 1
Scopus rating (2013): SJR 1.94 SNIP 1.197 CiteScore 3.52
ISI indexed (2013): ISI indexed yes
BFI (2012): BFI-level 1
Scopus rating (2012): SJR 1.94 SNIP 1.137 CiteScore 3.43
ISI indexed (2012): ISI indexed yes
BFI (2011): BFI-level 1
Scopus rating (2011): SJR 2.203 SNIP 1.223 CiteScore 3.73
ISI indexed (2011): ISI indexed yes
BFI (2010): BFI-level 1
Scopus rating (2010): SJR 2.737 SNIP 1.385
Web of Science (2010): Indexed yes
BFI (2009): BFI-level 1
Scopus rating (2009): SJR 2.759 SNIP 1.286
Web of Science (2009): Indexed yes
BFI (2008): BFI-level 2
Scopus rating (2008): SJR 2.426 SNIP 1.239
Web of Science (2008): Indexed yes
Scopus rating (2007): SJR 2.568 SNIP 1.238
Web of Science (2007): Indexed yes
Scopus rating (2006): SJR 2.554 SNIP 1.12
Scopus rating (2005): SJR 2.257 SNIP 0.969
Web of Science (2005): Indexed yes
Scopus rating (2004): SJR 1.51 SNIP 0.637
Scopus rating (2003): SJR 1.252 SNIP 0.531
Scopus rating (2002): SJR 1.245 SNIP 0.419
Higher variability in the number of sexual partners in males can contribute to a higher prevalence of sexually transmitted diseases in females

By examining published, empirical data we show that men and women consistently differ in the shape of the distribution of the number of sexual partners. The female distribution is always relatively narrow-variance is low—with a big majority of women having a number of partners close to the average. The male distribution is much wider-variance is high—with many men having few sex partners and many others having more partners than most females. Using stochastic modelling we demonstrate that this difference in variance is, in principle, sufficient to cause a difference in the gender prevalence of sexually transmitted diseases: compared to the situation where the genders have identical sex partner distributions, men will reach a lower equilibrium value, while women will stay at the same level (meaning that female prevalence becomes higher than male). We carefully analyse model behaviour and derive approximate expressions for equilibrium prevalences in the two different scenarios. We find that the size of the difference in gender prevalence depends on the variance ratio (the ratio between the variances of the male and female sex partner distributions), on the expected number of life-time partners, and on the probability of disease transmission. We note that in addition to humans, the variance phenomenon described here is likely to play a role for sexually transmitted diseases in other species also. We also show, again by examining published, empirical data, that the female to male prevalence ratio increases with the overall prevalence of a sexually transmitted disease (i.e., the more widespread the disease, the more women are affected). We suggest that this pattern may be caused by the effect described above in highly prevalent sexually transmitted diseases, while its impact in low-prevalence epidemics is surpassed by the action of high-risk individuals (mostly males).
InterMap3D: predicting and visualizing co-evolving protein residues

InterMap3D predicts co-evolving protein residues and plots them on the 3D protein structure. Starting with a single protein sequence, InterMap3D automatically finds a set of homologous sequences, generates an alignment and fetches the most similar 3D structure from the Protein Data Bank (PDB). It can also accept a user-generated alignment. Based on the alignment, co-evolving residues are then predicted using three different methods: Row and Column Weighing of Mutual Information, Mutual Information/Entropy and Dependency. Finally, InterMap3D generates high-quality images of the protein with the predicted co-evolving residues highlighted.

General information
State: Published
Organisations: Department of Systems Biology, Center for Biological sequence analysis, Center for Biological Sequence Analysis
Pages: 1963-1965
Publication date: 2009
Main Research Area: Technical/natural sciences
Contributions to the study of coevolution, at micro and macroscales

State: Published
Epitope mapping and analysis of sequence variation in VAR2CSA DBL3X involved in P-falciparum placental sequestration

Publication information
Place of publication: Kgs. Lyngby, Denmark
Publisher: Technical University of Denmark (DTU)
Original language: English
Main Research Area: Technical/natural sciences
Source: orbit
Source-ID: 260106
Publication: Research › Ph.D. thesis – Annual report year: 2008

General information
State: Published
Organisations: Center for Biological Sequence Analysis, Department of Systems Biology
Authors: Rask, T. S. (Intern), Dahlback, M. (Ekstern), Andersen, P. H. (Ekstern), Nielsen, M. A. (Ekstern), Ndam, N. T. (Ekstern), Resende, M. (Ekstern), Turner, L. (Ekstern), Deloron, P. (Ekstern), Hvid, L. (Ekstern), Lund, O. (Intern), Pedersen, A. G. (Intern), Theander, T. G. (Ekstern), Salanti, A. (Ekstern)
Pages: S17-S18
Publication date: 2008
Main Research Area: Technical/natural sciences

Publication information
Journal: Infection Genetics and Evolution
Volume: 8
Issue number: 4
ISSN (Print): 1567-1348
Ratings:
BFI (2018): BFI-level 1
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 1
Web of Science (2017): Indexed Yes
BFI (2016): BFI-level 1
Scopus rating (2016): CiteScore 2.77 SJR 1.306 SNIP 1.04
BFI (2015): BFI-level 1
Scopus rating (2015): SJR 1.435 SNIP 1.083 CiteScore 2.85
BFI (2014): BFI-level 1
Scopus rating (2014): SJR 1.376 SNIP 1.141 CiteScore 2.99
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 1
Scopus rating (2013): SJR 1.545 SNIP 1.174 CiteScore 3.26
ISI indexed (2013): ISI indexed yes
BFI (2012): BFI-level 1
Scopus rating (2012): SJR 1.216 SNIP 0.993 CiteScore 2.87
ISI indexed (2012): ISI indexed yes
BFI (2011): BFI-level 1
Scopus rating (2011): SJR 1.182 SNIP 1.194 CiteScore 3.11
ISI indexed (2011): ISI indexed yes
Web of Science (2011): Indexed yes
BFI (2010): BFI-level 1
Scopus rating (2010): SJR 1.313 SNIP 1.05
Web of Science (2010): Indexed yes
BFI (2009): BFI-level 1
Scopus rating (2009): SJR 1.263 SNIP 1.094
BFI (2008): BFI-level 1
Scopus rating (2008): SJR 1.092 SNIP 0.905
Plasmodium falciparum erythrocyte membrane protein 1 diversity analysis and classification.

General information
State: Published
Organisations: Center for Biological Sequence Analysis, Department of Systems Biology
Authors: Rask, T. S. (Intern), Lavstsen, T. (Ekstern), Pedersen, A. G. (Intern)
Pages: S98-S98
Publication date: 2008
Main Research Area: Technical/natural sciences

Publication information
Journal: International Journal For Parasitology
Volume: 38
ISSN (Print): 0020-7519
Ratings:
BFI (2018): BFI-level 1
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 1
Web of Science (2017): Indexed Yes
BFI (2016): BFI-level 1
Scopus rating (2016): SJR 1.845 SNIP 1.676 CiteScore 4.17
BFI (2015): BFI-level 1
Scopus rating (2015): SJR 2.038 SNIP 1.521 CiteScore 3.89
BFI (2014): BFI-level 1
Scopus rating (2014): SJR 1.789 SNIP 1.36 CiteScore 3.57
BFI (2013): BFI-level 1
Scopus rating (2013): SJR 1.774 SNIP 1.341 CiteScore 3.55
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
BFI (2012): BFI-level 1
Scopus rating (2012): SJR 1.542 SNIP 1.31 CiteScore 3.76
ISI indexed (2012): ISI indexed yes
BFI (2011): BFI-level 1
Scopus rating (2011): SJR 1.623 SNIP 1.427 CiteScore 3.8
ISI indexed (2011): ISI indexed yes
BFI (2010): BFI-level 1
Scopus rating (2010): SJR 1.645 SNIP 1.335
BFI (2009): BFI-level 1
Scopus rating (2009): SJR 1.636 SNIP 1.236
Web of Science (2009): Indexed yes
BFI (2008): BFI-level 1
Scopus rating (2008): SJR 1.816 SNIP 1.454
Web of Science (2008): Indexed yes
A human phenome-interactome network of protein complexes implicated in genetic disorders

We performed a systematic, large-scale analysis of human protein complexes comprising gene products implicated in many different categories of human disease to create a phenome-interactome network. This was done by integrating quality-controlled interactions of human proteins with a validated, computationally derived phenotype similarity score, permitting identification of previously unknown complexes likely to be associated with disease. Using a phenomic ranking of protein complexes linked to human disease, we developed a Bayesian predictor that in 298 of 669 linkage intervals correctly ranks the known disease-causing protein as the top candidate, and in 870 intervals with no identified disease-causing gene, provides novel candidates implicated in disorders such as retinitis pigmentosa, epithelial ovarian cancer, inflammatory bowel disease, amyotrophic lateral sclerosis, Alzheimer disease, type 2 diabetes and coronary heart disease. Our publicly available draft of protein complexes associated with pathology comprises 506 complexes, which reveal functional relationships between disease-promoting genes that will inform future experimentation.

General information
State: Published
Organisations: Department of Systems Biology, Center for Biological Sequence Analysis
Authors: Hansen, K. L. (Intern), Karlberg, E. O. L. (Intern), Størling, Z. M. (Intern), Ólason, P. Í. (Intern), Pedersen, A. G. (Intern), Rigina, O. (Intern), Hinsby, A. M. (Ekstern), Tumer, Z. (Ekstern), Pociot, F. (Ekstern), Tommerup, N. (Ekstern), Moreau, Y. (Ekstern), Brunak, S. (Intern)
Pages: 309-16
Publication date: 2007
Main Research Area: Technical/natural sciences

Publication information
Journal: Nature Biotechnology
Volume: 25
Issue number: 3
ISSN (Print): 1087-0156
Ratings:
BFI (2018): BFI-level 3
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 2
Web of Science (2017): Indexed Yes
BFI (2016): BFI-level 2
Scopus rating (2016): CiteScore 13.16 SJR 20.253 SNIP 6.303
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 2
Scopus rating (2015): SJR 17.892 SNIP 5.505 CiteScore 11.88
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 2
Scopus rating (2014): SJR 16.443 SNIP 5.433 CiteScore 11.4
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 2
Scopus rating (2013): SJR 13.849 SNIP 5.416 CiteScore 10.45
Evidence for Vertical Inheritance and Loss of the Leukotoxin Operon in Genus Mannheimia

The Mannheimia subclades belong to the same bacterial genus but have taken divergent paths toward their distinct lifestyles. M. haemolytica + M. glucosida are potential pathogens of the respiratory tract in the mammalian suborder Ruminantia, whereas M. ruminalis, the supposed sister group, lives as a commensal in the ovine rumen. We have tested the hypothesis that horizontal gene transfer of the leukotoxin operon has catalyzed pathogenic adaptation and speciation of M. haemolytica + M. glucosida, or other major subclades, by using a strategy that combines compositional and phylogenetic methods. We show that it has been vertically inherited from the last common ancestor of the diverging Mannheimia subclades, although several strains belonging to M. ruminalis have lost the operon. Our analyses support that divergence within M. ruminalis following colonization of the ovine rumen was very rapid and that functional decay of most of the leukotoxin operons occurred early when the adaptation to the rumen was fastest, suggesting that antagonistic pleiotropy was the main contributor to losses in the radiating lineages of M. ruminalis. To sum up, the scenario derived from these analyses reflects two aspects. On one hand, it opposes the hypothesis of horizontal gene transfer as a catalyst of pathogenic adaptation and speciation. On the other hand, it indicates that losses of the leukotoxin operons in the radiating lineages of M. ruminalis have catalyzed their adaptation to a commensal environment and reproductive isolation (speciation).

General information

State: Published
Organisations: Center for Biological Sequence Analysis, Department of Systems Biology, Bacteriology & Pathology, Division of Veterinary Diagnostics and Research, National Veterinary Institute
ABSTRACT: BACKGROUND: Some amino acid residues functionally interact with each other. This interaction will result in an evolutionary co-variation between these residues - coevolution. Our goal is to find these coevolving residues.

RESULTS: We present six new methods for detecting coevolving residues. Among other things, we suggest measures that are variants of Mutual Information, and measures that use a multidimensional representation of each residue in order to capture the physico-chemical similarities between amino acids. We created a benchmarking system, in silico, able to evaluate these methods through a wide range of realistic conditions. Finally, we use the combination of different methods as a way of improving performance. CONCLUSION: Our best method (Row and Column Weighed Mutual Information) has an estimated accuracy increase of 63% over Mutual Information. Furthermore, we show that the combination of different methods is efficient, and that the methods are quite sensitive to the different conditions tested.
Identification of a novel Mannheimia granulomatis lineage from pathological lesions in roe deer (Capreolus capreolus)

Eight atypical Mannheimia isolates were isolated from lesions in roe deer (Capreolus capreolus). Traditional classification based on morphologic and physiologic traits showed that they belong to a distinct biogroup (taxon) within genus Mannheimia. Extensive phenotypic characterization suggested that the isolates should be classified as M. granulomatis, although the presence of distinct traits justified their classification into a separate biogroup within this species. Phylogenetic analyses based on 16S rRNA sequences from two roe deer isolates and 41 other Mannheimia strains supported that the roe deer isolates form a monophyletic group within M. granulomatis. The lktA genotype was present in all roe deer isolates based on Southern blot analysis, whereas the corresponding beta-hemolytic phenotype was absent in one of these isolates.

General information
State: Published
Organisations: Center for Biological Sequence Analysis, Department of Systems Biology
Authors: Bojesen, A. M. (Ekstern), Larsen, J. (Ekstern), Pedersen, A. G. (Intern), Mörner, T. (Ekstern), Mattson, R. (Ekstern), Bisgaard, M. (Ekstern)
Pages: 345-52
Publication date: 2007
Main Research Area: Technical/natural sciences

Publication information
Journal: Journal of wildlife diseases
Issue number: 3
ISSN (Print): 0090-3558
Ratings:
BFI (2018): BFI-level 1
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 1
Web of Science (2017): Indexed Yes
BFI (2016): BFI-level 1
Scopus rating (2016): CiteScore 1.52 SJR 0.787 SNIP 0.922
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 1
Scopus rating (2015): SJR 0.757 SNIP 0.841 CiteScore 1.34
BFI (2014): BFI-level 1
Scopus rating (2014): SJR 0.829 SNIP 0.797 CiteScore 1.37
BFI (2013): BFI-level 1
Scopus rating (2013): SJR 0.763 SNIP 0.843 CiteScore 1.45
ISI indexed (2013): ISI indexed yes
BFI (2012): BFI-level 1
Scopus rating (2012): SJR 0.738 SNIP 0.941 CiteScore 1.51
ISI indexed (2012): ISI indexed yes
Web of Science (2012): Indexed yes
BFI (2011): BFI-level 1
Scopus rating (2011): SJR 0.602 SNIP 0.695 CiteScore 1.26
ISI indexed (2011): ISI indexed yes
Web of Science (2011): Indexed yes
BFI (2010): BFI-level 1
Scopus rating (2010): SJR 0.819 SNIP 0.907
BFI (2009): BFI-level 1
Scopus rating (2009): SJR 0.766 SNIP 0.967
BFI (2008): BFI-level 1
Scopus rating (2008): SJR 0.747 SNIP 0.936
Scopus rating (2007): SJR 0.554 SNIP 0.793
Web of Science (2007): Indexed yes
Scopus rating (2006): SJR 0.596 SNIP 0.958
Scopus rating (2005): SJR 0.536 SNIP 0.847
In order to shed light on the nature of the persistent reservoir of human immunodeficiency virus type 1 (HIV-1), we investigated signs of recent evolution in the pool of proviral DNA in patients on successful HAART. Pro-viral DNA, corresponding to the C2-V3-C3 region of the HIV-1 env gene, was collected from PBMCs isolated from 57 patients. Both "consensus" (57 patients) and clonal (7 patients) sequences were obtained from five time points spanning a 24-month period. The main computational strategy was to use maximum likelihood to fit a set of alternative phylogenetic models to the clonal data, and then determine the support for models that imply evolution between time points. Model fit and model-selection uncertainty was assessed using the Akaike information criterion (AIC) and Akaike weights. The consensus sequence data was also analyzed using a range of phylogenetic techniques to determine whether there were temporal trends indicating ongoing replication and evolution. In summary, it was not possible to detect definitive signs of ongoing evolution in either the bulk-sequenced or the clonal data with the methods employed here, but our results could be consistent with localized expression of archival HIV genomes in some patients. Interestingly, stop-codons were present at the same two positions in several independent clones and across patients. Simulation studies indicated that this phenomenon could be explained as the result of parallel evolution and that some sites were inherently more likely to evolve into stop codons.

**General information**

**State:** Published  
**Organisations:** Center for Biological Sequence Analysis, Department of Systems Biology  
**Authors:** Mens, H. (Ekstern), Pedersen, A. G. (Intern), Jørgensen, L. B. (Ekstern), Hue, S. (Ekstern), Gerstoft, J. (Ekstern), Katzenstein, T. L. (Ekstern)  
**Pages:** 107-115  
**Publication date:** 2007  
**Main Research Area:** Technical/natural sciences

**Publication information**

**Journal:** AIDS research and human retroviruses  
**Issue number:** 1  
**ISSN (Print):** 0889-2229  
**Ratings:**  
- BFI (2018): BFI-level 1  
- Web of Science (2018): Indexed yes  
- BFI (2017): BFI-level 1  
- Web of Science (2017): Indexed Yes  
- BFI (2016): BFI-level 1  
- Scopus rating (2016): SJR 1.039 SNIP 0.572 CiteScore 1.66  
- BFI (2015): BFI-level 1  
- Scopus rating (2015): SJR 1.037 SNIP 0.583 CiteScore 1.67  
- BFI (2014): BFI-level 1  
- Scopus rating (2014): SJR 1.147 SNIP 0.648 CiteScore 2.01  
- BFI (2013): BFI-level 1  
- Scopus rating (2013): SJR 1.124 SNIP 0.715 CiteScore 2.15  
- ISI indexed (2013): ISI indexed yes  
- Web of Science (2013): Indexed yes  
- BFI (2012): BFI-level 1  
- Scopus rating (2012): SJR 1.147 SNIP 0.783 CiteScore 2.34
MaxAlign: maximizing usable data in an alignment

BACKGROUND: The presence of gaps in an alignment of nucleotide or protein sequences is often an inconvenience for bioinformatical studies. In phylogenetic and other analyses, for instance, gapped columns are often discarded entirely from the alignment. RESULTS: MaxAlign is a program that optimizes the alignment prior to such analyses. Specifically, it maximizes the number of nucleotide (or amino acid) symbols that are present in gap-free columns in the alignment area. By selecting the optimal subset of sequences to exclude from the alignment, MaxAlign can be used prior to phylogenetic and bioinformatical analyses as well as in other situations where this form of alignment improvement is useful. In this work, we test MaxAlign's performance in these tasks and compare the accuracy of phylogenetic estimates including and excluding gapped columns from the analysis, with and without processing with MaxAlign. In this paper, we also introduce a new simple measure of tree similarity, Normalized Symmetric Similarity (NSS) that we consider useful for comparing tree topologies. CONCLUSION: We demonstrate how MaxAlign is helpful in detecting misaligned or defective sequences without requiring manual inspection. We also show that it is not advisable to exclude gapped columns from phylogenetic analyses unless MaxAlign is used first. Finally, we find that the sequences removed by MaxAlign from an alignment tend to be those that would otherwise be associated with low phylogenetic accuracy, and that the presence of gaps in any given sequence does not seem to disturb the phylogenetic estimates of other sequences. The MaxAlign web-server is freely available online at http://www.cbs.dtu.dk/services/MaxAlign where supplementary information can also be found. The program is also freely available as a Perl stand-alone package.
Epitope mapping and topographic analysis of VAR2CSA DBL3X involved in P.falciparum placental sequestration

Pregnancy-associated malaria is a major health problem, which mainly affects primigravidae living in malaria endemic areas. The syndrome is precipitated by accumulation of infected erythrocytes in placental tissue through an interaction between chondroitin sulphate A on syncytiotrophoblasts and a parasite-encoded protein on the surface of infected erythrocytes, believed to be VAR2CSA. VAR2CSA is a polymorphic protein of approximately 3,000 amino acids forming six Duffy-binding-like (DBL) domains. For vaccine development it is important to define the antigenic targets for protective antibodies and to characterize the consequences of sequence variation. In this study, we used a combination of in silico tools, peptide arrays, and structural modeling to show that sequence variation mainly occurs in regions under strong diversifying selection, predicted to form flexible loops. These regions are the main targets of naturally acquired immunoglobulin gamma and accessible for antibodies reacting with native VAR2CSA on infected erythrocytes. Interestingly, surface reactive anti-VAR2CSA antibodies also target a conserved DBL3X region predicted to form an alpha-helix. Finally, we could identify DBL3X sequence motifs that were more likely to occur in parasites isolated from primi- and multigravidae, respectively. These findings strengthen the vaccine candidacy of VAR2CSA and will be important for choosing epitopes and variants of DBL3X to be included in a vaccine protecting women against pregnancy-associated malaria.

General information
State: Published
Organisations: Center for Biological Sequence Analysis, Department of Systems Biology
Authors: Dahlback, M. (Ekstern), Rask, T. S. (Intern), Andersen, P. (Intern), Nielsen, M. A. (Ekstern), Ndam, N. T. (Ekstern), Resende, M. (Ekstern), Turner, L. (Ekstern), Deloron, P. (Ekstern), Hvid, L. (Ekstern), Lund, O. (Intern), Pedersen, A. G. (Intern), Theander, T. G. (Ekstern), Salanti, A. (Ekstern)
Pages: 1069-1082
Publication date: 2006
Main Research Area: Technical/natural sciences

Publication information
Journal: PLoS Pathogens
Volume: 2
Issue number: 11
ISSN (Print): 1553-7366
Ratings:
BFI (2018): BFI-level 2
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 2
Web of Science (2017): Indexed Yes
BFI (2016): BFI-level 2
Scopus rating (2016): SJR 4.466 SNIP 1.635 CiteScore 6.46
BFI (2015): BFI-level 2
Scopus rating (2015): SJR 5.019 SNIP 1.783 CiteScore 7.14
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 2
Scopus rating (2014): SJR 5.227 SNIP 1.926 CiteScore 7.67
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 2
Scopus rating (2013): SJR 5.26 SNIP 1.924 CiteScore 8.22
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
BFI (2012): BFI-level 2
Scopus rating (2012): SJR 4.825 SNIP 1.845 CiteScore 8.33
ISI indexed (2012): ISI indexed yes
Web of Science (2012): Indexed yes
BFI (2011): BFI-level 2
Scopus rating (2011): SJR 5.442 SNIP 1.933 CiteScore 8.87
ISI indexed (2011): ISI indexed yes
BFI (2010): BFI-level 2
Scopus rating (2010): SJR 4.494 SNIP 1.702
Web of Science (2010): Indexed yes
BFI (2009): BFI-level 2
Scopus rating (2009): SJR 3.424 SNIP 1.435
Identification of Tn5397-like and Tn916-like transposons and diversity of the tetracycline resistance gene tet(M) in enterococci from humans, pigs and poultry

General information
State: Published
Organisations: Division of Microbiology and Risk Assessment, National Food Institute, Center for Biological Sequence Analysis, Department of Systems Biology
Authors: Agersø, Y. (Intern), Pedersen, A. G. (Intern), Aarestrup, F. M. (Intern)
Pages: 832-839
Publication date: 2006
Main Research Area: Technical/natural sciences

Publication information
Journal: JOURNAL OF ANTIMICROBIAL CHEMOTHERAPY
Volume: 57
Issue number: 5
ISSN (Print): 0305-7453
Ratings:
BFI (2018): BFI-level 1
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 1
Web of Science (2017): Indexed yes
BFI (2016): BFI-level 1
Scopus rating (2016): CiteScore 4.21 SJR 2.24 SNIP 1.527
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 1
Scopus rating (2015): SJR 2.203 SNIP 1.513 CiteScore 4.06
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 1
Scopus rating (2014): SJR 2.303 SNIP 1.772 CiteScore 4.61
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 1
Scopus rating (2013): SJR 2.416 SNIP 1.782 CiteScore 4.7
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
BFI (2012): BFI-level 1
Scopus rating (2012): SJR 2.157 SNIP 1.654 CiteScore 4.35
ISI indexed (2012): ISI indexed yes
Web of Science (2012): Indexed yes
BFI (2011): BFI-level 1
Scopus rating (2011): SJR 2.235 SNIP 1.745 CiteScore 4.24
Protein evolution is faster outside the cell

General information
State: Published
Organisations: Center for Biological sequence analysis, Center for Biological Sequence Analysis, Department of Systems Biology
Authors: Julenius, K. (Intern), Pedersen, A. G. (Intern)
Pages: 2039-2048
Publication date: 2006
Main Research Area: Technical/natural sciences

Publication information
Journal: Molecular Biology and Evolution
Volume: 23
ISSN (Print): 0737-4038

Ratings:
BFI (2018): BFI-level 2
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 2
Web of Science (2017): Indexed yes
BFI (2016): BFI-level 2
Scopus rating (2016): SJR 8.724 SNIP 7.289 CiteScore 13.93
Web of Science (2016): Indexed yes
Characterization of rotavirus strains in a Danish population: high frequency of mixed infections and diversity within the VP4 gene of \( P \) [8] strains

**General information**

**State:** Published

**Organisations:** Center for Biological Sequence Analysis, Department of Systems Biology

**Authors:** Fischer, T. (Ekstern), Eugen-Olsen, J. (Ekstern), Pedersen, A. G. (Intern), Molbak, K. (Ekstern), Bottiger, B. (Ekstern), Rostgaard, K. (Ekstern), Nielsen, N. (Ekstern)

**Pages:** 1099-1104

**Publication date:** 2005

**Main Research Area:** Technical/natural sciences

**Publication information**

**Journal:** JOURNAL OF CLINICAL MICROBIOLOGY

**Volume:** 43

**Original language:** English

**Links:**


**Source:** orbit
Genetic evolution of HIV in patients remaining on a stable HAART regimen despite insufficient viral suppression

General information
State: Published
Organisations: Department of Systems Biology
Authors: Kristiansen, T. (Ekstern), Pedersen, A. G. (Intern), Eugen-Olsen, J. (Ekstern), Katzenstein, T. (Ekstern)
Pages: 890-901
Publication date: 2005
Main Research Area: Technical/natural sciences

Publication information
Journal: SCANDINAVIAN JOURNAL OF INFECTIOUS DISEASES
Volume: 37
Issue number: 11-12
ISSN (Print): 0036-5548
Ratings:
BFI (2018): BFI-level 1
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 1
Web of Science (2017): Indexed Yes
BFI (2016): BFI-level 1
Scopus rating (2016): SJR 0.718 SNIP 0.637
BFI (2015): BFI-level 1
Scopus rating (2015): SJR 0.84 SNIP 0.661 CiteScore 1.36
BFI (2014): BFI-level 1
Scopus rating (2014): SJR 0.766 SNIP 0.71 CiteScore 1.61
BFI (2013): BFI-level 1
Scopus rating (2013): SJR 0.823 SNIP 0.841 CiteScore 1.71
ISI indexed (2013): ISI indexed yes
BFI (2012): BFI-level 1
Scopus rating (2012): SJR 0.756 SNIP 0.778 CiteScore 1.66
ISI indexed (2012): ISI indexed yes
BFI (2011): BFI-level 1
Scopus rating (2011): SJR 0.751 SNIP 0.726 CiteScore 1.63
ISI indexed (2011): ISI indexed yes
BFI (2010): BFI-level 1
Scopus rating (2010): SJR 0.71 SNIP 0.775
BFI (2009): BFI-level 1
Scopus rating (2009): SJR 0.625 SNIP 0.678
BFI (2008): BFI-level 1
Scopus rating (2008): SJR 0.624 SNIP 0.76
Scopus rating (2007): SJR 0.581 SNIP 0.792
Web of Science (2007): Indexed yes
Scopus rating (2006): SJR 0.691 SNIP 0.793
Scopus rating (2005): SJR 0.599 SNIP 0.713
Web of Science (2005): Indexed yes
Scopus rating (2004): SJR 0.499 SNIP 0.646
Scopus rating (2003): SJR 0.489 SNIP 0.734
Scopus rating (2002): SJR 0.555 SNIP 0.657
Web of Science (2002): Indexed yes
Scopus rating (2001): SJR 0.576 SNIP 0.719
Scopus rating (2000): SJR 0.585 SNIP 0.819
Scopus rating (1999): SJR 0.562 SNIP 0.813
Limited inter- and intra-patient sequence diversity of the genetic lineage a human metapneumovirus fusion gene

Human metapneumovirus (hMPV) is associated with respiratory tract illness especially in young children. Two hMPV genetic lineages, A and B, and four sublineages A1, A2 and B1, B2 have been defined. Infection with hMPV occurs through membrane fusion mediated by the hMPV fusion (F) protein. In this study, the inter- and intra-patient genetic diversity of the lineage A hMPV F gene was investigated. Ten isolates were collected from 10 hMPV infected children. Viral RNA was isolated and amplified, and approximately 10 clones from each isolate were sequenced. Altogether 108 clones were successfully sequenced. The average interpatient sequence diversity was 1.68% and 1.64% at nucleotide and amino acid levels, respectively. The samples were divisible into two groups on the basis of intrapatient sequence diversity. In group 1 (4 children) the intra-patient sequence diversity was low (nt: 0.26-0.39%, aa: 0.51-0.94%) whereas group 2 (6 children) had a higher intra-patient sequence diversity (nt: 0.85-1.98%, aa: 1.08-2.22%). Phylogenetic analyses showed that the group 1 children harboured sublineage A1 only, but interestingly group 2 children harboured both sublineages A1 and A2, indicating they had been infected with at least two viruses. Several independent viruses contained premature stop codons in exactly identical positions resulting in truncated fusion proteins. Possibly this is a mechanism for immune system evasion. The F protein is a major antigenic determinant, and the limited sequence diversity observed lay emphasis on the hMPV F gene as a putative target for future vaccine development.

General information
State: Published
Organisations: Center for Biological Sequence Analysis, Department of Systems Biology
Authors: Winther, T. (Ekstern), Madsen, C. (Ekstern), Pedersen, A. G. (Intern), Von Linstow, M. (Ekstern), Eugen-Olsen, J. (Ekstern), Hogh, B. (Ekstern)
Pages: 89-97
Publication date: 2005
Main Research Area: Technical/natural sciences

Publication information
Journal: Virus Genes
Volume: 31
Issue number: 1
ISSN (Print): 0920-8569
Ratings:
BFI (2018): BFI-level 1
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 1
Web of Science (2017): Indexed Yes
BFI (2016): BFI-level 1
Scopus rating (2016): SJR 0.635 SNIP 0.685 CiteScore 1.55
BFI (2015): BFI-level 1
Scopus rating (2015): SJR 0.759 SNIP 0.773 CiteScore 1.55
BFI (2014): BFI-level 1
Scopus rating (2014): SJR 0.729 SNIP 0.774 CiteScore 1.62
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 1
Scopus rating (2013): SJR 0.84 SNIP 0.934 CiteScore 1.94
ISI indexed (2013): ISI indexed yes
BFI (2012): BFI-level 1
Scopus rating (2012): SJR 0.755 SNIP 0.809 CiteScore 1.8
ISI indexed (2012): ISI indexed yes
BFI (2011): BFI-level 1
Scopus rating (2011): SJR 0.834 SNIP 0.996 CiteScore 1.92
ISI indexed (2011): ISI indexed yes
BFI (2010): BFI-level 1
Scopus rating (2010): SJR 0.742 SNIP 0.852
BFI (2009): BFI-level 1
Major histocompatibility complex (MHC) proteins are encoded by extremely polymorphic genes and play a crucial role in immunity. However, not all genetically different MHC molecules are functionally different. Sette and Sidney (1999) have defined nine HLA class I supertypes and showed that with only nine main functional binding specificities it is possible to cover the binding properties of almost all known HLA class I molecules. Here we present a comprehensive study of the functional relationship between all HLA molecules with known specificities in a uniform and automated way. We have developed a novel method for clustering sequence motifs. We construct hidden Markov models for HLA class I molecules using a Gibbs sampling procedure and use the similarities among these to define clusters of specificities. These clusters are extensions of the previously suggested ones. We suggest splitting some of the alleles in the A1 supertype into a new A26 supertype, and some of the alleles in the B27 supertype into a new B39 supertype. Furthermore, the B8 alleles may define their own supertype. We also use the published specificities for a number of HLA-DR types to define clusters with similar specificities. We report that the previously observed specificities of these class II molecules can be clustered into nine classes, which only partly correspond to the serological classification. We show that classification of HLA molecules may be done in a uniform and automated way. The definition of clusters allows for selection of representative HLA molecules that can cover the HLA specificity space better. This makes it possible to target most of the known HLA alleles with known specificities using only a few peptides, and may be used in construction of vaccines. Supplementary material is available at http://www.cbs.dtu.dk/researchgroups/immunology/supertypes.html.
Characterization of incompletely typed rotavirus strains from Guinea-Bissau: identification of G8 and G9 types and a high frequency of mixed infections

Among 167 rotavirus specimens collected from young children in a suburban area of Bissau, Guinea-Bissau, from 1996 to 1998, most identifiable strains belonged to the uncommon P[6], G2 type and approximately 50% remained incompletely typed. In the present study, 76 such strains were further characterized. Due to interprimer interaction during the standard multiplex PCR approach, modifications of this procedure were implemented. The modified analyses revealed a high frequency of G2, G8, and G9 genotypes, often combined with P[4] and/or P[6]. The Guinean G8 and G9 strains were 97
and 98%, respectively, identical to other African G8 and G9 strains. Multiple G and/or P types were identified at a high frequency (59%), including two previously undescribed mixed infections, P[4]P[6], G2G8 and P[4]P[6], G2G9. These mixed infections most likely represent naturally occurring reassortance of rotavirus strains. Detection of such strains among the previously incompletely typed strains indicates a potential underestimation of mixed infections, if only a standard multiplex PCR procedure is followed. Furthermore cross-priming of the G3 primer with the G8 primer binding site and silent mutations at the P[4] and P[6] primer binding sites were detected. These findings highlight the need for regular evaluation of the multiplex primer PCR method and typing primers. The high frequency of uncommon as well as reassortant rotavirus strains in countries where rotavirus is an important cause of child mortality underscores the need for extensive strain surveillance as a basis to develop appropriate rotavirus vaccine candidates.
RevTrans: multiple alignment of coding DNA from aligned amino acid sequences

The simple fact that proteins are built from 20 amino acids while DNA only contains four different bases, means that the 'signal-to-noise ratio' in protein sequence alignments is much better than in alignments of DNA. Besides this information-theoretical advantage, protein alignments also benefit from the information that is implicit in empirical substitution matrices such as BLOSUM-62. Taken together with the generally higher rate of synonymous mutations over non-synonymous ones, this means that the phylogenetic signal disappears much more rapidly from DNA sequences than from the encoded proteins. It is therefore preferable to align coding DNA at the amino acid level and it is for this purpose we have constructed the program RevTrans. RevTrans constructs a multiple DNA alignment by: (i) translating the DNA; (ii) aligning the resulting peptide sequences; and (iii) building a multiple DNA alignment by 'reverse translation' of the aligned protein sequences. In the resulting DNA alignment, gaps occur in groups of three corresponding to entire codons, and analogous codon positions are therefore always lined up. These features are useful when constructing multiple DNA alignments for phylogenetic analysis. RevTrans also accepts user-provided protein alignments for greater control of the alignment process. The RevTrans web server is freely available at http://www.cbs.dtu.dk/services/RevTrans/.
Flexibility of the genetic code with respect to DNA structure

Motivation. The primary function of DNA is to carry genetic information through the genetic code. DNA, however, contains a variety of other signals related, for instance, to reading frame, codon bias, pairwise codon bias, splice sites and transcription regulation, nucleosome positioning and DNA structure. Here we study the relationship between the genetic code and DNA structure and address two questions. First, to which degree does the degeneracy of the genetic code and the acceptable amino acid substitution patterns allow for the superimposition of DNA structural signals to protein coding sequences? Second, is the origin or evolution of the genetic code likely to have been constrained by DNA structure?

Results. We develop an index for code flexibility with respect to DNA structure. Using five different di- or tri-nucleotide models of sequence-dependent DNA structure, we show that the standard genetic code provides a fair level of flexibility at the level of broad amino acid categories. Thus the code generally allows for the superimposition of any structural signal on any protein-coding sequence, through amino acid substitution. The flexibility observed at the level of single amino acids allows only for the superimposition of punctual and loosely positioned signals to conserved amino acid sequences. The degree of flexibility of the genetic code is low or average with respect to several classes of alternative codes. This result is consistent with the view that DNA structure is not likely to have played a significant role in the origin and evolution of the genetic code.

General information

State: Published
Organisations: Department of Systems Biology, Center for Biological Sequence Analysis, University of California, Irvine
Authors: Baisnée, P. F. (Ekstern), Baldi, P. (Ekstern), Brunak, S. (Intern), Pedersen, A. G. (Intern)
Number of pages: 12
Pages: 237-248
A DNA structural atlas for *Escherichia coli*

**General information**

State: Published

Organisations: Department of Biotechnology

Authors: Pedersen, A. G. (Intern), Jensen, L. J. (Ekstern), Brunak, S. (Intern), Stærfeldt, H. H. (Intern), Ussery, D. (Intern)

Pages: 907-930

Publication date: 2000

Main Research Area: Technical/natural sciences

**Publication information**

Journal: Journal of Molecular Biology

Volume: 299

Issue number: 4

ISSN (Print): 0022-2836

Ratings:

BFI (2018): BFI-level 1

Web of Science (2018): Indexed yes

BFI (2017): BFI-level 1

Web of Science (2017): Indexed Yes

BFI (2016): BFI-level 1

Scopus rating (2016): CiteScore 4.13 SJR 3.265 SNIP 1.156

BFI (2015): BFI-level 1

Scopus rating (2015): SJR 2.979 SNIP 1.105 CiteScore 3.97

Web of Science (2015): Indexed yes

BFI (2014): BFI-level 1

Scopus rating (2014): SJR 2.806 SNIP 1.117 CiteScore 3.7

BFI (2013): BFI-level 1

Scopus rating (2013): SJR 3.112 SNIP 1.095 CiteScore 3.92

ISI indexed (2013): ISI indexed yes

BFI (2012): BFI-level 1

Scopus rating (2012): SJR 3.01 SNIP 1.139 CiteScore 3.91

ISI indexed (2012): ISI indexed yes

BFI (2011): BFI-level 1

Scopus rating (2011): SJR 3.092 SNIP 1.192 CiteScore 4.01

ISI indexed (2011): ISI indexed yes

Web of Science (2011): Indexed yes

BFI (2010): BFI-level 1

Scopus rating (2010): SJR 3.11 SNIP 1.16

Web of Science (2010): Indexed yes

BFI (2009): BFI-level 1

Scopus rating (2009): SJR 3.45 SNIP 1.159

Web of Science (2009): Indexed yes

BFI (2008): BFI-level 2

Scopus rating (2008): SJR 3.612 SNIP 1.206

Web of Science (2008): Indexed yes

Scopus rating (2007): SJR 3.803 SNIP 1.292

Web of Science (2007): Indexed yes

Scopus rating (2006): SJR 3.887 SNIP 1.332

Web of Science (2006): Indexed yes
Structural basis for triplet repeat disorders: A computational analysis

Motivation: Over a dozen major degenerative disorders, including myotonic dystrophy, Huntington's disease and fragile X syndrome result from unstable expansions of particular trinucleotides. Remarkably, only some of all the possible triplets, namely CAG/CTG, CGG/CCG and GAA/TTC, have been associated with the known pathological expansions. This raises some basic questions at the DNA level. Why do particular triplets seem to be singled out? What is the mechanism for their expansion and how does it depend on the triplet itself? Could other triplets or longer repeats be involved in other diseases?

Results: Using several different computational models of DNA structure, we show that the triplets involved in the pathological repeats generally fall into extreme classes. Thus, CAG/CTG repeats are particularly flexible, whereas GCC, CGG and GAA repeats appear to display both flexible and rigid (but curved) characteristics depending on the method of analysis. The fact that (1) trinucleotide repeats often become increasingly unstable when they exceed a length of approximately 50 repeats, and (2) repeated 12-mers display a similar increase in instability above 13 repeats, together suggest that approximately 150 bp is a general threshold length for repeat instability. Since this is about the length of DNA wrapped up in a single nucleosome care particle, we speculate that chromatin structure may play an important role in the expansion mechanism. We furthermore suggest that expansion of a dodecamer repeat, which we predict to have very high flexibility, may play a role in the pathogenesis of the neurodegenerative disorder multiple system atrophy (MSA).

General information
State: Published
Organisations: Department of Biotechnology, University of California, Irvine, Net-ID Inc.
Authors: Baldi, P. (Ekstern), Brunak, S. (Intern), Chauvin, Y. (Ekstern), Pedersen, A. G. (Intern)
Pages: 918-929
Publication date: 1999
Main Research Area: Technical/natural sciences

Publication information
Journal: Bioinformatics
Volume: 15
Issue number: 11
ISSN (Print): 1367-4803
Ratings:
BFI (2018): BFI-level 2
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 2
Web of Science (2017): Indexed yes
BFI (2016): BFI-level 2
Scopus rating (2016): CiteScore 6.42
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 2
Scopus rating (2015): CiteScore 6.06
Web of Science (2015): Indexed yes
Computational prediction of eukaryotic promoters from the nucleotide sequence is one of the most attractive problems in sequence analysis today, but it is also a very difficult one. Thus, current methods predict in the order of one promoter per kilobase in human DNA, while the average distance between functional promoters has been estimated to be in the range of 30-40 kilobases. Although it is conceivable that some of these predicted promoters correspond to cryptic initiation sites that are used in vivo, it is likely that most are false positives. This suggests that it is important to carefully reconsider the biological data that forms the basis of current algorithms, and we here present a review of data that may be useful in this regard. The review covers the following topics: (1) basal transcription and core promoters, (2) activated transcription and transcription factor binding sites, (3) CpG islands and DNA methylation, (4) chromosomal structure and nucleosome modification, and (5) chromosomal domains and domain boundaries. We discuss the possible lessons that may be learned, especially with respect to the wealth of information about epigenetic regulation of transcription that has been appearing in recent years. (C) 1999 Elsevier Science Ltd. All rights reserved.
Computational analyses and annotations of the Arabidopsis peroxidase gene family

General information
State: Published
Organisations: Department of Biotechnology, University of Copenhagen
Authors: Østergaard, L. (Ekstern), Pedersen, A. G. (Intern), Jespersen, H. M. (Ekstern), Brunak, S. (Intern), Welinder, K. G. (Ekstern)
Pages: 98-102
Publication date: 1998
Main Research Area: Technical/natural sciences

Publication information
Journal: F E B S Letters
Volume: 433
ISSN (Print): 0014-5793
Ratings:
BFI (2018): BFI-level 1
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 1
Web of Science (2017): Indexed Yes
BFI (2016): BFI-level 1
Scopus rating (2016): CiteScore 3.48 SJR 1.898 SNIP 0.885
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 1
Scopus rating (2015): SJR 2.02 SNIP 0.927 CiteScore 3.49
BFI (2014): BFI-level 1
Scopus rating (2014): SJR 1.86 SNIP 0.871 CiteScore 3.19
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 1
Scopus rating (2013): SJR 2.328 SNIP 0.984 CiteScore 3.71
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
BFI (2012): BFI-level 1
Scopus rating (2012): SJR 2.259 SNIP 0.914 CiteScore 3.67
ISI indexed (2012): ISI indexed yes
Web of Science (2012): Indexed yes
BFI (2011): BFI-level 1
Scopus rating (2011): SJR 2.264 SNIP 0.837 CiteScore 3.5
ISI indexed (2011): ISI indexed yes
Web of Science (2011): Indexed yes
BFI (2010): BFI-level 1
Scopus rating (2010): SJR 2.197 SNIP 0.795
Web of Science (2010): Indexed yes
Computational applications of DNA physical scales
The authors study from a computational standpoint several different physical scales associated with structural features of DNA sequences, including dinucleotide scales such as base stacking energy and propeller twist, and trinucleotide scales such as bendability and nucleosome positioning. We show that these scales provide an alternative or complementary compact representation of DNA sequences. As an example we construct a strand invariant representation of DNA sequences. The scales can also be used to analyze and discover new DNA structural patterns, especially in combination with hidden Markov models (HMMs). The scales are applied to HMMs of human promoter sequences revealing a number of significant differences between regions upstream and downstream of the transcriptional start point. Finally we show, with some qualifications, that such scales are by and large independent, and therefore complement each other.

General information
State: Published
Organisations: Department of Biotechnology, Net-ID Inc.
Authors: Baldi, P. (Ekstern), Chauvin, Y. (Ekstern), Brunak, S. (Intern), Gorodkin, J. (Intern), Pedersen, A. G. (Intern)
Pages: 35-42
Publication date: 1998

Host publication information
Title of host publication: Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology
Place of publication: Menlo Park
Publisher: AAAI Press
Main Research Area: Technical/natural sciences
Conference: Sixth International Conference on Intelligent Systems for Molecular Biology, Montreal, Canada, 01/01/1998
Source: orbit
Source-ID: 170925
Publication: Research - peer-review › Article in proceedings – Annual report year: 1998

Computational applications of DNA structural scales
Studies several different physical scales associated with the structural features of DNA sequences from a computational standpoint, including dinucleotide scales, such as base stacking energy and propeller twist, and trinucleotide scales, such as bendability and nucleosome positioning. We show that these scales provide an alternative or complementary compact representation of DNA sequences. As an example, we construct a strand-invariant representation of DNA sequences. The scales can also be used to analyze and discover new DNA structural patterns, especially in combination with hidden
Markov models (HMMs). The scales are applied to HMMs of human promoter sequences, revealing a number of significant differences between regions upstream and downstream of the transcriptional start-point. Finally, we show (with some qualifications) that such scales are, by and large, independent, and therefore complement each other.

**General information**

State: Published  
Organisations: Department of Systems Biology, Center for Biological Sequence Analysis, Department of Biotechnology, Net-ID Inc.  
Authors: Baldi, P. (Ekstern), Chauvin, Y. (Ekstern), Brunak, S. (Intern), Gorodkin, J. (Intern), Pedersen, A. G. (Intern)  
Number of pages: 8  
Publication date: 1998

**Host publication information**

Title of host publication: ISMB-98 Proceedings  
Publisher: AAAI Press  
Main Research Area: Technical/natural sciences  
Conference: Sixth International Conference on Intelligent Systems for Molecular Biology, Montreal, Canada, 01/01/1998  
Source: Findit  
Source-ID: 23541236  
Publication: Research - peer-review › Article in proceedings – Annual report year: 1998

**DNA structure in human RNA polymerase II promoters**

**General information**

State: Published  
Organisations: Department of Biotechnology, Net-ID Inc.  
Authors: Pedersen, A. G. (Intern), Baldi, P. (Ekstern), Chauvin, Y. (Ekstern), Brunak, S. (Intern)  
Pages: 663-673  
Publication date: 1998  
Main Research Area: Technical/natural sciences

**Publication information**

Journal: Journal of Molecular Biology  
Volume: 281  
ISSN (Print): 0022-2836  
Ratings:  
BFI (2018): BFI-level 1  
Web of Science (2018): Indexed yes  
BFI (2017): BFI-level 1  
Web of Science (2017): Indexed Yes  
BFI (2016): BFI-level 1  
Scopus rating (2016): CiteScore 4.13 SJR 3.265 SNIP 1.156  
BFI (2015): BFI-level 1  
Scopus rating (2015): SJR 2.979 SNIP 1.105 CiteScore 3.97  
Web of Science (2015): Indexed yes  
BFI (2014): BFI-level 1  
Scopus rating (2014): SJR 2.806 SNIP 1.117 CiteScore 3.7  
BFI (2013): BFI-level 1  
Scopus rating (2013): SJR 3.112 SNIP 1.095 CiteScore 3.92  
ISI indexed (2013): ISI indexed yes  
BFI (2012): BFI-level 1  
Scopus rating (2012): SJR 3.01 SNIP 1.139 CiteScore 3.91  
ISI indexed (2012): ISI indexed yes  
BFI (2011): BFI-level 1  
Scopus rating (2011): SJR 3.092 SNIP 1.192 CiteScore 4.01  
ISI indexed (2011): ISI indexed yes  
Web of Science (2011): Indexed yes  
BFI (2010): BFI-level 1  
Scopus rating (2010): SJR 3.11 SNIP 1.16  
Web of Science (2010): Indexed yes
Neural Network Prediction of Translation Initiation Sites in Eukaryotes: Perspectives for EST and Genome analysis

Translation in eukaryotes does not always start at the first AUG in an mRNA, implying that context information also plays a role. This makes prediction of translation initiation sites a non-trivial task, especially when analysing EST and genome data where the entire mature mRNA sequence is not known. In this paper, we employ artificial neural networks to predict which AUG triplet in an mRNA sequence is the start codon. The trained networks correctly classified 88% of Arabidopsis and 85% of vertebrate AUG triplets. We find that our trained neural networks use a combination of local start codon context and global sequence information. Furthermore, analysis of false predictions shows that AUGs in frame with the actual start codon are more frequently selected than out-of-frame AUGs, suggesting that our networks use reading frame detection. A number of conflicts between neural network predictions and database annotations are analysed in detail, leading to identification of possible database errors.

General information
State: Published
Organisations: Center for Biological sequence analysis, Department of Systems Biology
Authors: Pedersen, A. G. (Intern), Nielsen, H. (Intern)
Pages: 226-233
Publication date: 1997

Host publication information
Title of host publication: Proceedings Fifth International Conference on Intelligent Systems for Molecular Biology
Place of publication: Menlo Park, CA, USA
Publisher: AAAI Press
Main Research Area: Technical/natural sciences
Conference: 5th International Conference on Intelligent Systems for Molecular Biology, Halkidiki, Greece, 21/06/1997 - 21/06/1997
Source: orbit
Source-ID: 168458
Publication: Research - peer-review » Article in proceedings – Annual report year: 1997

Characterization of prokaryotic and eukaryotic promoters using hidden Markov models
In this paper we utilize hidden Markov models (HMMs) and information theory to analyze prokaryotic and eukaryotic promoters. We perform this analysis with special emphasis on the fact that promoters are divided into a number of
different classes, depending on which polymerase-associated factors that bind to them. We find that HMMs trained on such subclasses of Escherichia coli promoters (specifically, the so-called sigma 70 and sigma 54 classes) give an excellent classification of unknown promoters with respect to sigma-class. HMMs trained on eukaryotic sequences from human genes also model nicely all the essential well known signals, in addition to a potentially new signal upstream of the TATA-box. We furthermore employ a novel technique for automatically discovering different classes in the input data (the promoters) using a system of self-organizing parallel HMMs. These self-organizing HMMs have at the same time the ability to find clusters and the ability to model the sequential structure in the input data. This is highly relevant in situations where the variance in the data is high, as is the case for the subclass structure in for example promoter sequences.

General information
State: Published
Organisations: Department of Systems Biology, Center for Biological Sequence Analysis, Net-ID Inc., California Institute of Technology
Authors: Pedersen, A. G. (Intern), Baldi, P. (Ekster), Chauvin, Y. (Ekster), Brunak, S. (Intern)
Number of pages: 10
Pages: 182-191
Publication date: 1996

Characterization of prokaryotic and eukaryotic promoters using hidden Markov models
In this paper we utilize hidden Markov models (HMMs) and information theory to analyze prokaryotic and eukaryotic promoters. We perform this analysis with special emphasis on the fact that promoters are divided into a number of different classes, depending on which polymerase-associated factors that bind to them. We find that HMMs trained on such subclasses of Escherichia coli promoters (specifically, the so-called sigma-70 and sigma-54 classes) give an excellent classification of unknown promoters with respect to sigma-class. HMMs trained on eukaryotic sequences from human genes also model nicely all the essential well known signals, in addition to a potentially new signal upstream of the TATA-box. We furthermore employ a novel technique for automatically discovering different classes in the input data (the promoters) using a system of self-organizing parallel HMMs. These self-organizing HMMs have at the same time the ability to find clusters and the ability to model the sequential structure in the input data. This is highly relevant in situations where the variance in the data is high, as is the case for the subclass structure in for example promoter sequences.

General information
State: Published
Organisations: Department of Chemistry, California Institute of Technology, Net-ID Inc.
Authors: Pedersen, A. G. (Intern), Baldi, P. (Ekster), Brunak, S. (Intern), Chauvin, Y. (Ekster)
Pages: 182-191
Publication date: 1996
We present a novel method for using the learning ability of a neural network as a measure of information in local regions of input data. Using the method to analyze Escherichia coli promoters, we discover all previously described signals, and furthermore find new signals that are regularly spaced along the promoter region. The spacing of all signals correspond to the helical periodicity of DNA, meaning that the signals are all present on the same face of the DNA helix in the promoter region. This is consistent with a model where the RNA polymerase contacts the promoter on one side of the DNA, and suggests that the regions important for promoter recognition may include more positions on the DNA than usually assumed. We furthermore analyze the E. coli promoters by calculating the Kullback Leibler distance, and by constructing sequence logos.
Systems Biology of the Infant Gut Microbiome

Department of Systems Biology
Period: 01/10/2017 → 30/09/2020
Number of participants: 4
Phd Student: Myers, Pernille Neve (Intern)
Supervisor: Nielsen, Henrik Bjørn (Intern)
Pedersen, Anders Gorm (Intern)
Main Supervisor: Pedersen, Susanne Brix (Intern)

Financing sources

Biomarkers for prognosis and prediction of childhood ALL treatment outcome

Department of Bio and Health Informatics
Period: 01/11/2016 → 31/10/2019
Number of participants: 5
Phd Student: Nielsen, Rikke Linnemann (Intern)
Supervisor: Pedersen, Anders Gorm (Intern)
Schmiegelow, Kjeld (Ekstern)
Wang, XiuJie (Ekstern)
Main Supervisor: Gupta, Ramneek (Intern)

Financing sources

Investigation og the genetic basis for virus tropism and virulence of classical swine fever virus

National Veterinary Institute
Period: 15/08/2015 → 14/08/2018
Number of participants: 4
Phd Student: Johnston, Camille Melissa (Intern)
Supervisor: Belsham, Graham (Intern)
Pedersen, Anders Gorm (Intern)
Main Supervisor: Rasmussen, Thomas Bruun (Intern)

Financing sources

Understanding aetiology and treatment trajectories in childhood leukemia through advanced data integration

Source: Internal funding (public)
Name of research programme: Samfinansieret - Andet
Project: PhD
Genus-level studies of gene dynamics for the Aspergillus genus

Department of Systems Biology
Period: 01/12/2014 → 30/01/2018
Number of participants: 5
Phd Student:
Theobald, Sebastian (Intern)
Supervisor:
Larsen, Thomas Ostenfeld (Intern)
Pedersen, Anders Gorm (Intern)
Vesth, Tammi Camilla (Intern)
Main Supervisor:
Andersen, Mikael Rørdam (Intern)

Financing sources
Source: Internal funding (public)
Name of research programme: Samfinansierede - Virksomhed
Project: PhD

Genus-level studies of genome dynamics for the Aspergillus genus

Department of Systems Biology
Period: 01/11/2014 → 31/12/2017
Number of participants: 4
Phd Student:
Rasmussen, Jane Lind Nybo (Intern)
Supervisor:
Pedersen, Anders Gorm (Intern)
Vesth, Tammi Camilla (Intern)
Main Supervisor:
Andersen, Mikael Rørdam (Intern)

Financing sources
Source: Internal funding (public)
Name of research programme: Samfinansierede - Virksomhed
Project: PhD

Identification of risk factors for acquiring ADV in Danish mink farms

Department of Bio and Health Informatics
Period: 01/08/2013 → 06/06/2017
Number of participants: 7
Phd Student:
Hagberg, Emma Elisabeth (Intern)
Human Personality - Identification of important genotypic and phenotypic factors for the development of an individual's personality

Department of Systems Biology
Period: 15/12/2012 → 29/02/2016
Number of participants: 6
Phd Student:
Wolffhechel, Karin Marie Brandt (Intern)
Supervisor:
Pedersen, Anders Gorm (Intern)
Main Supervisor:
Jarmer, Hanne Østergaard (Intern)
Examiner:
Nielsen, Morten (Intern)
Lindgren, Cecilia Margareta (Ekstern)
Reuter, Martin (Ekstern)

Financing sources
Source: Internal funding (public)
Name of research programme: Institut stipendie (DTU)
Project: PhD

Microbial Community Interactions in Arctic Environments using a Metagenomics Approach

Department of Bio and Health Informatics
Period: 15/12/2012 → 26/04/2017
Number of participants: 6
Phd Student:
Hauptmann, Aviaja Zenia Edna Lyberth (Intern)
Main Supervisor:
Sicheritz-Pontén, Thomas (Intern)
Examiner:
Pedersen, Anders Gorm (Intern)
Winding, Anne (Ekstern)
Øvreås, Lise (Ekstern)

Financing sources
Source: Internal funding (public)
Name of research programme: Institut stipendie (DTU) Samf.

Relations
Publications:
Microbial Biogeography of the Arctic Cryosphere
Project: PhD
A systems biology approach for type 2 diabetes patient stratification using multiple date types

Department of Systems Biology
Period: 01/07/2012 → 15/12/2015
Number of participants: 6
Phd Student:
Gudmundsdottir, Valborg (Intern)
Supervisor:
Gupta, Ramneek (Intern)
Main Supervisor:
Brunak, Søren (Intern)
Examiner:
Pedersen, Anders Gorm (Intern)
Jensen, Thomas Skøt (Intern)
Zeggini, Eleftheria (Ekstern)

Financing sources
Source: Internal funding (public)
Name of research programme: 1/3 FUU, 1/3 inst 1/3 Andet
Project: PhD

Targeting the genetic complexity within adapting RNA virus populations

National Veterinary Institute
Period: 15/12/2011 → 27/05/2015
Number of participants: 6
Phd Student:
Fahnøe, Ulrik (Intern)
Supervisor:
Pedersen, Anders Gorm (Intern)
Main Supervisor:
Rasmussen, Thomas Bruun (Intern)
Examiner:
Polacek, Charlotta (Intern)
Ruggli, Nicolas (Ekstern)
Vignuzzi, Marco (Ekstern)

Financing sources
Source: Internal funding (public)
Name of research programme: Institut stipendie (DTU) Samf.
Project: PhD

Regulation of RNA expression in solid tumors

Department of Systems Biology
Period: 01/11/2011 → 30/09/2015
Number of participants: 6
Phd Student:
Favero, Francesco (Intern)
Supervisor:
Eklund, Aron Charles (Intern)
Main Supervisor:
Szallasi, Zoltan Imre (Intern)
Examiner:
Pedersen, Anders Gorm (Intern)
Csabai, István (Ekstern)
Lamy, Philippe (Ekstern)

Financing sources
Phylogenomics approaches for enzyme discovery

Department of Systems Biology
Period: 01/10/2011 → 30/06/2016
Number of participants: 6
Phd Student: Özen, Asli Ismihan (Intern)
Supervisor: Pedersen, Anders Gorm (Intern)
Main Supervisor: Sicheritz-Pontén, Thomas (Intern)
Examiner: Petersen, Bent (Intern)
Ahrén, Dag Gustaf (Ekstern)
Pisani, Davide (Ekstern)

Financing sources
Source: Internal funding (public)
Name of research programme: Institut stipendie (DTU) Samf.
Project: PhD

Evolution and Co-evolution of Systemic Regulatory Mechanisms

Department of Systems Biology
Period: 01/01/2010 → 30/04/2011
Number of participants: 3
Phd Student: Jacobsen, Janus Valentin (Intern)
Supervisor: Jensen, Lars Juhl (Intern)
Main Supervisor: Pedersen, Anders Gorm (Intern)

Financing sources
Source: Internal funding (public)
Name of research programme: Eksternt finansieret virksomhed
Project: PhD

Genetic Variation and human disease

Department of Systems Biology
Period: 01/10/2009 → 01/07/2015
Number of participants: 6
Phd Student: Nielsen, Kasper (Intern)
Supervisor: Gupta, Ramneek (Intern)
Main Supervisor: Brunak, Søren (Intern)
Examiner: Pedersen, Anders Gorm (Intern)
Tolstrup, Niels (Intern)
Warinner, Christina (Ekstern)

Financing sources
Source: Internal funding (public)
Name of research programme: Institut stipendie (DTU) Samf.
**Malarian Evolution og Vaccine Design**

Department of Systems Biology  
Period: 01/11/2007 → 18/04/2012  
Number of participants: 5  
Phd Student: Hansen, Daniel Aaen (Intern)  
Main Supervisor: Pedersen, Anders Gorm (Intern)  
Examiner: Lundegaard, Claus (Intern)  
Salanti, Ali (Ekstern)  
Smith, Joseph Douglas (Ekstern)

**Financing sources**  
Source: Internal funding (public)  
Name of research programme: DTU-lønnet stipendie  
Project: PhD

**Human monoclonal antibody-assisted characterization of Plasmodium falciparum variant surface antigens**

Department of Systems Biology  
Period: 01/01/2007 → 21/12/2010  
Number of participants: 7  
Phd Student: Sørli, Jorid Birkelund (Intern)  
Supervisor: Barfod, Lea (Ekstern)  
Hviid, Lars (Ekstern)  
Main Supervisor: Lund, Ole (Intern)  
Examiner: Pedersen, Anders Gorm (Intern)  
Christensen, Jan Pravsgaard (Ekstern)  
Deloron, Philippe (Ekstern)

**Financing sources**  
Source: Internal funding (public)  
Name of research programme: 1/3 DTU-stip, 2/3 FUR/andet  
Project: PhD

**Biosimulation - A New Tool In Drug Development**

Department of Systems Biology  
Period: 01/07/2006 → 02/03/2011  
Number of participants: 4  
Phd Student: Li, Qiyuan (Intern)  
Main Supervisor: Szallasi, Zoltan Imre (Intern)  
Examiner: Pedersen, Anders Gorm (Intern)  
Wiuf, Carsten (Ekstern)

**Financing sources**  
Source: Internal funding (public)  
Name of research programme: 1/3 DTU-stip, 2/3 FUR/andet  
Project: PhD
Systems Biology of Pre-mRNA Splicing and its Role in the Development of Cancer

Department of Systems Biology
Period: 01/03/2006 → 21/10/2009
Number of participants: 6
Phd Student:
Wang, Kai (Intern)
Supervisor:
Wernersson, Rasmus (Intern)
Main Supervisor:
Brunak, Søren (Intern)
Examiner:
Pedersen, Anders Gorm (Intern)
Gorodkin, Jan (Intern)
Rouzé, Pierre (Ekstern)

Financing sources
Source: Internal funding (public)
Name of research programme: Anden EU-finansiering
Project: PhD

Darwinian Vaccine-Development

Department of Systems Biology
Period: 15/08/2005 → 21/04/2010
Number of participants: 6
Phd Student:
Rask, Thomas Salhøj (Intern)
Supervisor:
Lund, Ole (Intern)
Main Supervisor:
Pedersen, Anders Gorm (Intern)
Examiner:
Workman, Christopher (Intern)
Jensen, Anja T. R. (Ekstern)
Kesmir, Can (Intern)

Financing sources
Source: Internal funding (public)
Name of research programme: Forskningsrådsfinansiering
Project: PhD

Creation of a database of dangerous organisms

Department of Systems Biology
Number of participants: 6
Phd Student:
Pletscher-Frankild, Sune (Intern)
Supervisor:
Lundegaard, Claus (Intern)
Main Supervisor:
Lund, Ole (Intern)
Examiner:
Pedersen, Anders Gorm (Intern)
Borghans, José A.M. (Ekstern)
Frekiær, Hanne (Intern)

Financing sources
Data Mining and data integration in biotechnology

Department of Systems Biology
Period: 01/10/2003 → 25/03/2008
Number of participants: 5
Phd Student:
Ólason, Páll Ísólfur (Intern)
Main Supervisor:
Brunak, Søren (Intern)
Examiner:
Pedersen, Anders Gorm (Intern)
Jensen, Ole Nørregaard (Ekstern)
Rice, Peter Martin (Ekstern)

Financing sources
Source: Internal funding (public)
Name of research programme: Eksternt finansieret virksomhed
Project: PhD

Predicting Toxicity from Microarray Data

Department of Systems Biology
Period: 01/10/2003 → 30/05/2008
Number of participants: 7
Phd Student:
Spicker, Jeppe (Intern)
Supervisor:
Nielsen, Henrik Bjørn (Intern)
Pedersen, Henrik Toft (Ekstern)
Main Supervisor:
Brunak, Søren (Intern)
Examiner:
Pedersen, Anders Gorm (Intern)
Fleckner, Jan (Ekstern)
Holmes, Elaine (Ekstern)

Financing sources
Source: Internal funding (public)
Name of research programme: Grundforskningsfonden
Project: PhD

Molecular Evolution of Human Immunodeficiency Virus (HIV) and Hepatitis C Virus (HCV)

Department of Systems Biology
Period: 01/09/2003 → 02/07/2008
Number of participants: 5
Phd Student:
Oliveira, Rodrigo Gouveia (Intern)
Main Supervisor:
Pedersen, Anders Gorm (Intern)
Examiner:
Nielsen, Henrik (Intern)
Hurst, Laurence D. (Ekstern)
Schierup, Mikkel (Ekstern)

Financing sources
Source: Internal funding (public)
Name of research programme: Privatist
Project: PhD

**Bioinformatics and Systems Biology of the Cell Cycle**
Department of Systems Biology
Period: 01/06/2002 → 30/09/2005
Number of participants: 5
Phd Student:
de Lichtenberg, Ulrik (Intern)
Main Supervisor:
Brunak, Søren (Intern)
Examiner:
Pedersen, Anders Gorm (Intern)
Bähler, Jürg (Ekstern)
Ideker, Trey (Ekstern)

**Financing sources**
Source: Internal funding (public)

Name of research programme: Grundforskningsfonden
Project: PhD

**Bioinformatics for Drug Target Discovery: Detection of Novel Ion Channel Coding Genes**
Department of Systems Biology
Period: 15/10/2001 → 03/06/2005
Number of participants: 6
Phd Student:
Kiemer, Lars (Intern)
Supervisor:
Mikkelsen, Jens D. (Ekstern)
Main Supervisor:
Brunak, Søren (Intern)
Examiner:
Pedersen, Anders Gorm (Intern)
Kjems, Jørgen (Ekstern)
Ouzounis, Christos A. (Ekstern)

**Financing sources**
Source: Internal funding (public)

Name of research programme: Samarbejdsaftalefinans
Project: PhD

**Genome Wide transcription analysis of s. cerevisiae and human cell**
Department of Systems Biology
Number of participants: 6
Phd Student:
Rundsten, Carsten Friis (Intern)
Supervisor:
Brunak, Søren (Intern)
Main Supervisor:
Knudsen, Steen (Intern)
Examiner:
Pedersen, Anders Gorm (Intern)
Clausen, Ib Groth (Ekstern)
Tolstrup, Niels (Intern)

**Financing sources**
Source: Internal funding (public)
**Regulation of Transcription in Bacillus Subtilis**

Department of Systems Biology  
**Period:** 01/02/1999 → 24/01/2003  
**Number of participants:** 8  
**PhD Student:** Jarmer, Hanne Østergaard (Intern)  
**Supervisor:** Brunak, Søren (Intern)  
Nielsen, Allan Kent (Intern)  
Saxild, Hans Henrik (Intern)  
**Main Supervisor:** Knudsen, Steen (Intern)  
**Examiner:** Pedersen, Anders Gorm (Intern)  
Bron, Sierd (Ekstern)  
Tolstrup, Niels (Intern)

**Financing sources**  
**Source:** Internal funding (public)  
**Name of research programme:** Forskningsrådsfinans

**Biological Sequence Analysis**

The Center for Biological Sequence Analysis conducts basic research in bioinformatics - a rapidly growing area in the field of molecular biology. The overall goal of the research is to obtain knowledge of the complex relations between sequence composition - the content and order of the chemical building blocks - and macromolecular structure and function. In addition, the research is aimed at creating entirely new possibilities in the study of evolutionary processes by comparison of sequence patterns across species. The tremendous growth in the amount of sequence and structure data has brought about a radical change in the possibilities of developing powerful computer methods for classification, prediction and comparison of molecular structure and function. The work of the center is especially focussed on novel data driven computational methods, such as artificial neural networks and hidden Markov models. Please see [http://www.cbs.dtu.dk/](http://www.cbs.dtu.dk/) for further information.

Department of Biotechnology  
**Period:** 01/01/1998 → 31/08/2003  
**Number of participants:** 21  
**Project participant:** Sorgenfrei Blom, Nikolaj (Intern)  
Boesen, Lone (Intern)  
Gorodkin, Jan (Intern)  
Hansen, Jan (Intern)  
Knudsen, Steen (Intern)  
Krogh, Anders Stærmorese (Intern)  
Pedersen, Anders Gorm (Intern)  
Rapacki, Kristoffer (Intern)  
Stærfeldt, Hans Henrik (Intern)  
Keiding, Johanne (Intern)  
Nielsen, Henrik (Intern)  
Dahlin, Kristine Bøje (Intern)  
Larsen, Thomas Schou (Intern)  
Gupta, Ramneek (Intern)  
Andersen, Claus A. (Intern)  
Worning, Peder (Intern)
Workman, Christopher (Intern)
Jarmer, Hanne Østergaard (Intern)
Ussery, David (Intern)
Jensen, Lars Juhl (Intern)
Project Manager, organisational: 
Brunak, Søren (Intern)

Financing sources
Source: Unknown
Name of research programme: Ukendt
Amount: 35,000,000.00 Danish Kroner
Source: Unknown
Name of research programme: Ukendt
Amount: 1,800,000.00 Danish Kroner

Computer-analyse af eukaryote DNA-sekvenser med betydning for transkription
Department of Systems Biology
Period: 01/10/1994 → 02/07/1998
Number of participants: 2
Phd Student:
Pedersen, Anders Gorm (Intern)
Main Supervisor:
Brunak, Søren (Intern)

Financing sources
Source: Internal funding (public)
Name of research programme: Grundforskningsfonden-SU
Project: PhD

Biological Sequence Analysis
The Center for Biological Sequence Analysis conducts basic research in bioinformatics - a rapidly growing area in the field of molecular biology. The overall goal of the research is to obtain knowledge of the complex relations between sequence composition - the content and order of the chemical building blocks - and macromolecular structure and function. In addition, the research is aimed at creating entirely new possibilities in the study of evolutionary processes by comparison of sequence patterns across species. The tremendous growth in the amount of sequence and structure data has brought about a radical change in the possibilities of developing powerful computer methods for classification, prediction and comparison of molecular structure and function. The work of the center is especially focussed on novel data driven computational methods, such as artificial neural networks and hidden Markov models.

Department of Chemistry
Department of Biotechnology
Period: 01/09/1993 → 31/12/1997
Number of participants: 20
Project participant:
Boesen, Lone (Intern)
Frimand, Kenneth (Intern)
Knudsen, Steen (Intern)
Krogh, Anders Stærmose (Intern)
Lund, Ole (Intern)
Pedersen, Anders Gorm (Intern)
Rapacki, Kristoffer (Intern)
Sorgenfrei Blom, Nikolaj (Intern)
Stærfeldt, Hans Henrik (Intern)
Bolshoy, Alexander (Intern)
Gorodkin, Jan (Intern)
Hansen, Jan (Intern)
Nielsen, Henrik (Intern)
Financing sources
Source: Unknown
Name of research programme: Ukendt
Amount: 25,000,000.00 Danish Kroner
Source: Unknown
Name of research programme: Ukendt
Amount: 250,000.00 Danish Kroner
Source: Unknown
Name of research programme: Ukendt
Amount: 325,000.00 Danish Kroner
Source: Unknown
Name of research programme: Ukendt
Amount: 860,000.00 Danish Kroner
Project