Identity By Descent

&

Hidden Markov Model

Jiang Chongyi

2016-05-05

Outline

- Introduction to IBD
- IBD application in autosomal recessive diseases
- Introduction to Hidden Markov model
- Hidden Markov model and IBD

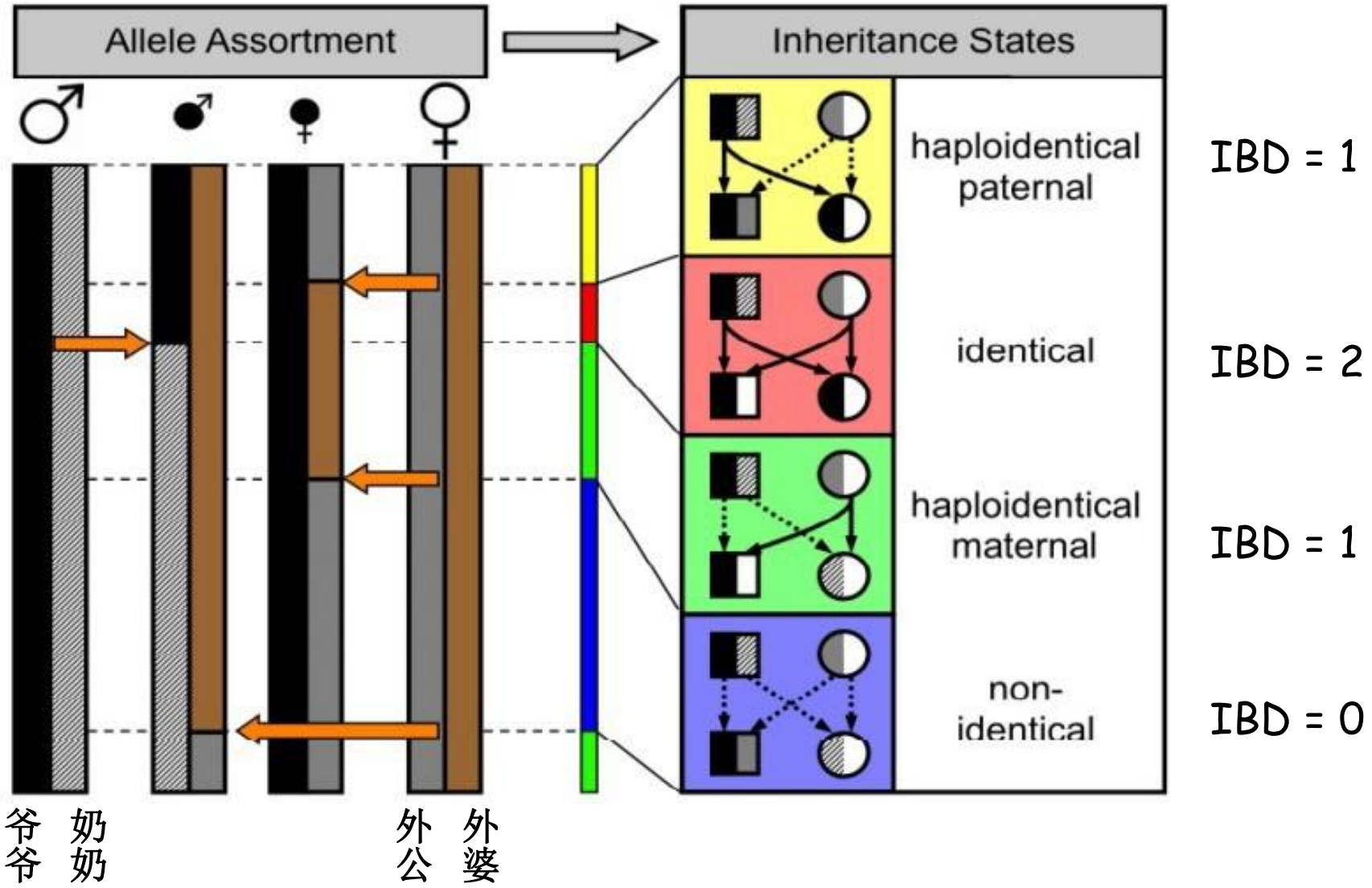
Introduction to IBD

IBS (identical by state)

vs.

IBD (identical by descent)

Introduction to IBD



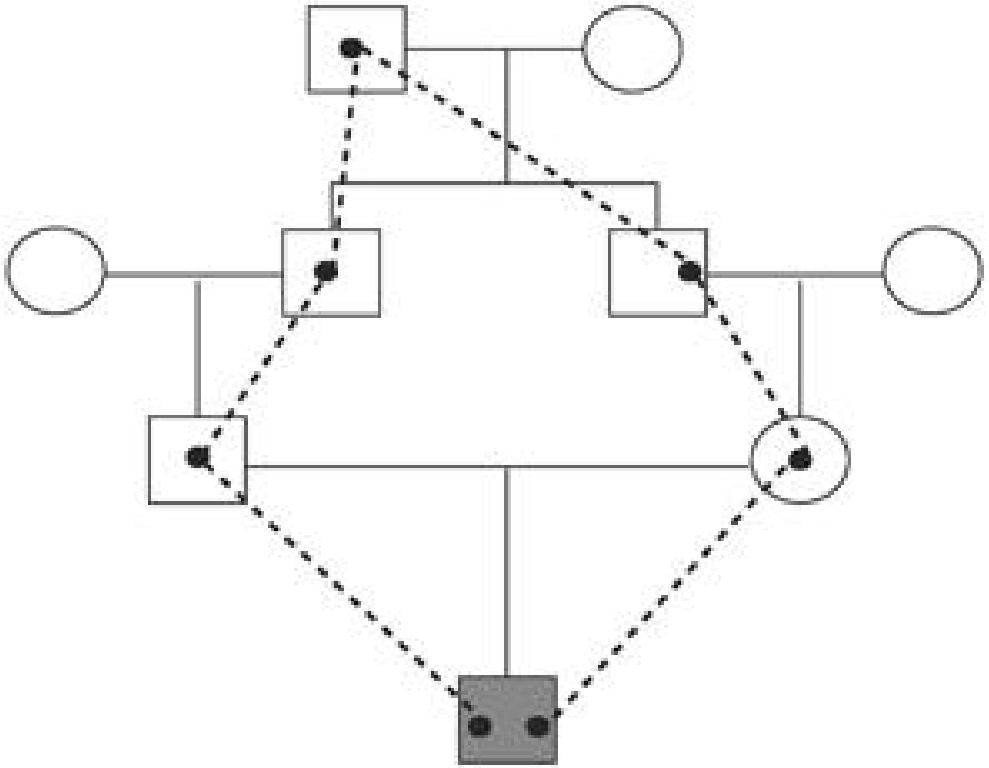
Roach JC, *et al.* Science (2010)

IBD application in autosomal recessive diseases

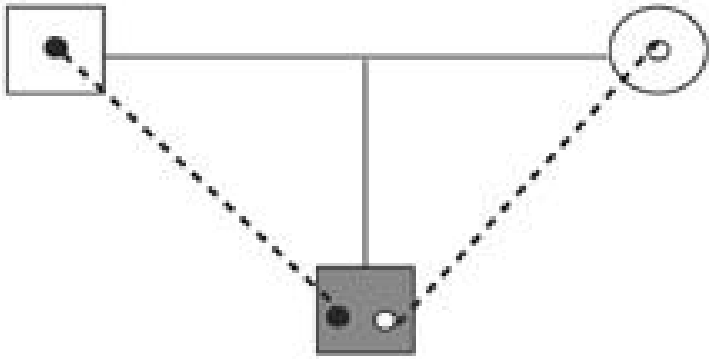
Application of IBD

- To quantify relatedness
- Genotype imputation and haplotype phase inference
- IBD in population genetics
- IBD mapping
- IBD application in autosomal recessive diseases

IBD application in autosomal recessive diseases



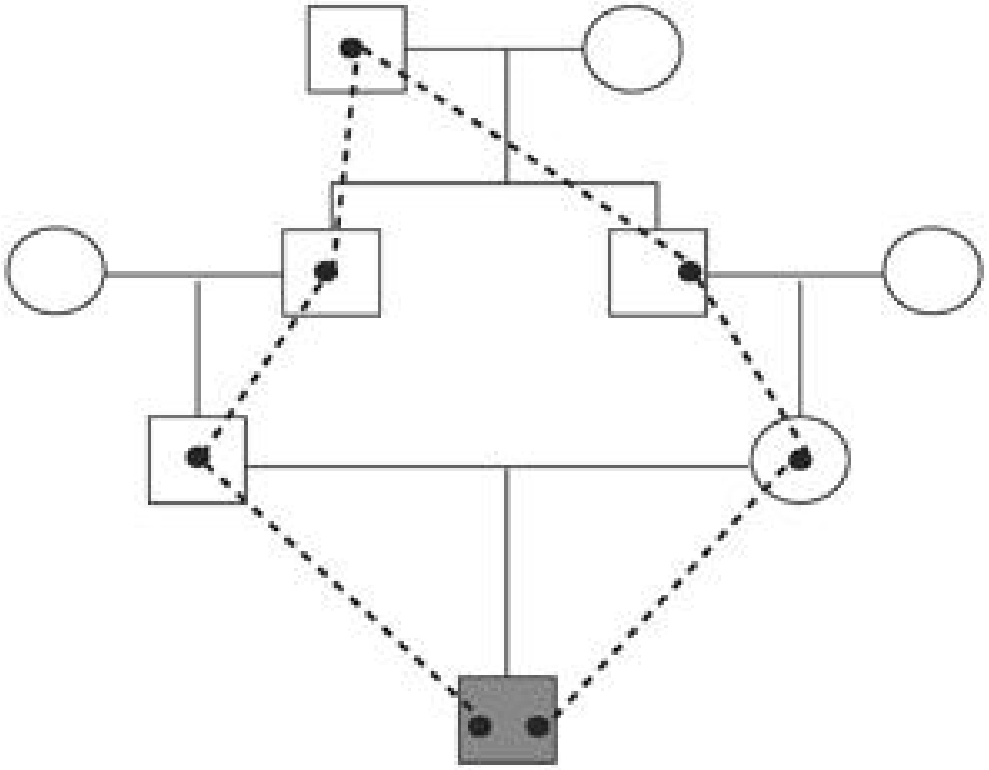
(a)



(b)

DOI: 10.1002/9780470015902.a0005407.pub2

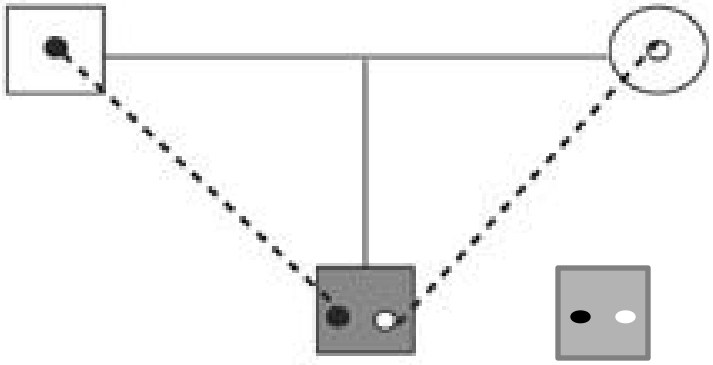
IBD application in autosomal recessive diseases



(a)

隐性纯合

Run of Homozygosity



(b)

复合杂合，隐性纯合

Identical by descent

IBD = 2

DOI: 10.1002/9780470015902.a0005407.pub2

IBD application in autosomal recessive diseases

BRIEF COMMUNICATIONS

nature genetics

Identity-by-descent filtering of exome sequence data identifies *PIGV* mutations in hyperphosphatasia mental retardation syndrome

Peter M Krawitz[1–3,11], Michal R Schweiger[1,2,11], Christian Rödelsperger[1–3], Carlo Marcelis[4], Uwe Kölsch[5], Christian Meisel[5], Friederike Stephani[5], Tarsh Kinoshita[6], Yoshiko Murakami[6], Sebastian Bauer[2], Melanie Isau[1], Axel Fischer[2], Andreas Dahl[2], Martin Kerick[2], Jochen Hecht[1,2], Sebastian Köhler[2], Marten Jäger[2], Johannes Grünhagen[2], Birgit Jonske de Condor[2], Sandra Doelken[2], Han G Brunner[4], Peter Meinecke[7], Eberhard Passarge[8], Miles D Thompson[9], David E Cole[9], Denise Horn[7], Tony Roscioli[6,10], Stefan Mundlos[1–3] & Peter N Robinson[1–3]

Hyperphosphatasia mental retardation (HPMR) syndrome is an autosomal recessive form of mental retardation with distinct facial features and elevated serum alkaline phosphatase. We performed whole-exome sequencing in three siblings of a nonconsanguineous union with HPMR and performed computational inference of regions identical by descent in all siblings to establish *PIGV*, encoding a member of the GPI-anchor biosynthesis pathway, as the gene mutated in HPMR. We identified homozygous or compound heterozygous mutations in *PIGV* in three additional families.

Recessive mutations are relatively common in the human genome, but their identification remains challenging. Initial efforts at using exome sequencing for disease gene discovery[1] analyzed small numbers of unrelated individuals, removed variants that are common or not predicted to be deleterious and then searched for genes with such variants in all affected individuals. The analysis of the exome sequences of two siblings and two further unrelated individuals affected by the autosomal recessive Miller syndrome led to the identification of *DHODH* as the disease gene[2]. Subsequently, researchers analyzed whole genome sequences of the same two siblings and their parents to identify chromosomal regions in which both siblings had inherited identical haplotypes from both parents, which allowed the number of gene candidates for Miller syndrome to be reduced from 34 to 4, showing that linkage information represents a useful filter for genome sequence data[3]. These studies illustrate the utility of sophisticated algorithmic analysis in reducing the candidate gene set beyond what can be achieved by a simple intersection filter.

HPMR, also known as Mabry syndrome (MIM%239300), was initially described as an autosomal recessive syndrome characterized by mental retardation and greatly elevated alkaline phosphatase levels[4,5]. Within a group of individuals with this rare syndrome, a previous study[6] delineated a specific clinical entity characterized by a distinct facial gestalt including hypertelorism, long palpebral fissures, a broad nasal bridge and tip, and a mouth with downturned corners and a thin upper lip, as well as brachytelephalangy. More variable neurological features included seizures and muscular hypotonia[6].

Here, DNA from three siblings of nonconsanguineous parents with this subtype of HPMR was analyzed by exome sequencing (**Supplementary Figs. 1** and **2** and **Supplementary Table 1**). Whole-exome sequencing using the ABI SOLiD platform was performed following enrichment of exonic sequences using Agilent's SureSelect whole-exome enrichment. Called variants were filtered to exclude variants not found in all affected persons as well as common variants identified in the dbSNP130 or HapMap databases, which left 14 candidate genes on multiple chromosomes (**Table 1** and **Supplementary Tables 2–4**).

In this work, we developed a statistical model that allowed us to infer regions that are identical by descent (IBD) from the exome sequences of only the affected children of a family in which an autosomal recessive disorder segregates. In consanguineous families, affected siblings share two haplotypes that are inherited from a single common ancestor at the disease locus and are thus homozygous by descent. In nonconsanguineous families, the affected children inherit identical maternal and paternal haplotypes in a region surrounding the disease gene, meaning that both haplotypes originated from the same maternal and paternal haplotype but are not necessarily from an identical ancestor (IBD = 2).

We developed an algorithm based on a Hidden Markov Model (HMM), a type of Bayesian network that is used to infer a sequence of hidden (that is, unobservable) states. We used the HMM algorithm to identify chromosomal regions with IBD = 2 in the presence of noisy (that is, potentially erroneous) sequence data. It is not possible to measure the IBD = 2 state directly; it is only possible to determine whether the genotypes of the siblings are compatible with identity-by-state status, that is, whether each sibling has the same homozygous

[1]Max Planck Institute for Molecular Genetics, Berlin, Germany. [2]Institut für Medizinische Genetik, Charité Universitätsmedizin Berlin, Berlin, Germany. [3]Berlin-Brandenburg Center for Regenerative Therapies (BCRT), Charité - Universitätsmedizin Berlin, Berlin, Germany. [4]Department of Human Genetics, University Medical Centre St. Radboud, Nijmegen, The Netherlands. [5]Institut für Medizinische Immunologie, Charité Universitätsmedizin Berlin, Berlin, Germany. [6]Department of Immunoregulation, Research Institute for Microbial Diseases, Osaka University, Osaka, Japan. [7]Medizinische Genetik, Altonaer Kinderkrankenhaus, Hamburg, Germany. [8]Institut für Humangenetik, Universitätsklinikum Essen, Essen, Germany. [9]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. [10]Department of Molecular and Clinical Genetics, University of Sydney, Sydney, Australia. [11]These authors contributed equally to this work. Correspondence should be addressed to S.M. (stefan.mundlos@charite.de) or P.N.R. (peter.robinson@charite.de).

Received 29 March; accepted 3 August; published online 29 August 2010; doi:10.1038/ng.652

BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 6 2011, pages 829–836 doi:10.1093/bioinformatics/btr022

Genetics and population analysis Advance Access publication January 28, 2011

Identity-by-descent filtering of exome sequence data for disease–gene identification in autosomal recessive disorders

Christian Rödelsperger[1,2,3,†], Peter Krawitz[1,2,3,†], Sebastian Bauer[2,†], Jochen Hecht[1,2,3], Abigail W. Bigham[4], Michael Bamshad[4], Birgit Jonske de Condor[1], Michal R. Schweiger[3] and Peter N. Robinson[1,2,3,*]

[1]Institute for Medical and Human Genetics, [2]Berlin-Brandenburg Center for Regenerative Therapies (BCRT), Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, Berlin, [3]Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany and [4]Department of Pediatrics, University of Washington, Seattle, WA 98195, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Next-generation sequencing and exome-capture technologies are currently revolutionizing the way geneticists screen for disease-causing mutations in rare Mendelian disorders. However, the identification of causal mutations is challenging due to the sheer number of variants that are identified in individual exomes. Although databases such as dbSNP or HapMap can be used to reduce the plethora of candidate genes by filtering out common variants, the remaining set of genes still remains on the order of dozens.

Results: Our algorithm uses a non-homogeneous hidden Markov model that employs local recombination rates to identify chromosomal regions that are identical by descent (IBD = 2) in children of consanguineous or non-consanguineous parents solely based on genotype data of siblings derived from high-throughput sequencing platforms. Using simulated and real exome sequence data, we show that our algorithm is able to reduce the search space for the causative disease gene to a fifth or a tenth of the entire exome.

Availability: An R script and an accompanying tutorial are available at http://compbio.charite.de/index.php/ibd2.html.

Contact: peter.robinson@charite.de

Received on October 9, 2010; revised on December 13, 2010; accepted on January 11, 2011

1 INTRODUCTION

The identification of genes underlying Mendelian disorders for the past several decades has mainly proceeded by means of positional cloning to identify chromosomal linkage intervals followed by the sequencing of candidate genes (Collins, 1995). Efforts at disease-gene identification involving linkage analysis or association studies usually result in a genomic interval of 0.5–10 cM containing up to 300 genes (Botstein and Risch, 2003). Although computational methods can be used to prioritize candidate genes (Köhler *et al.*, 2008), sequencing large numbers of candidate genes remains a time consuming and expensive task, and it is often not possible to identify the correct disease gene by inspection of the list of genes within the interval. Recently, whole-exome sequencing, i.e. the targeted capture of protein coding exons followed by massively parallel, 'next-generation' sequencing (NGS), has been demonstrated as an effective approach to identify genes underlying Mendelian disorders using a small number of affected individuals (Bieseckr, 2010).

Sequenced individuals typically have on the order of five to ten thousand variant calls representing either non-synonymous substitutions in protein coding sequences, alterations of the canonical splice-site dinucleotides or small indels (NS/SS/I) (Gilissen *et al.*, 2010; Ng *et al.*, 2009; Rios *et al.*, 2010). Even after filtering out common variants using data from dbSNP, the HapMap project and related resources such as the 1000 Genomes project, the number of potentially disease-causing NS/SS/I variants can remain high if the exome of a single patient is considered in isolation. Many disease-causing mutations were completely unsuspected on the basis of previous knowledge (Altshuler *et al.*, 2008), and software tools that aim at predicting the damaging effect of non-synonymous variants (Adzhubei *et al.*, 2010; Kumar *et al.*, 2009; Schwarz *et al.*, 2010; Sunyaev *et al.*, 2001) are currently unable to reliably distinguish between disease-causing mutations and other variants.

Groups who have performed disease–gene identification projects by exome sequencing (Choi *et al.*, 2009; Hoischen *et al.*, 2010; Ng *et al.*, 2009, 2010b) have developed analysis strategies based upon searching for potentially damaging rare variants found in the same gene in sets of multiple unrelated patients affected by the same Mendelian disorder. Although this strategy has been applied successfully in sequencing projects with two affected individuals (Gilissen *et al.*, 2010; Lalonde *et al.*, 2010) and occasionally even with a single affected individual (Pierce *et al.*, 2010; Rios *et al.*, 2010), in many cases multiple candidate genes remain after applying computational filters based on rarity or presence of a mutation in multiple affected patients (Hoischen *et al.*, 2010; Ng *et al.*, 2009, 2010a, b). This means that additional analysis of multiple candidate genes or other procedures would often be needed to identify the disease gene following exome sequencing of single families with a Mendelian disorder.

We will refer to the above-described procedure for searching for a disease gene by exome sequencing in multiple unrelated patients as

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

IBD application in autosomal recessive diseases

Hyperphosphatasia mental retardation (HPMR) syndrome

- Autosomal Recessive
- Mental retardation
- Distinct facial features
- Elevated serum alkaline phosphatase

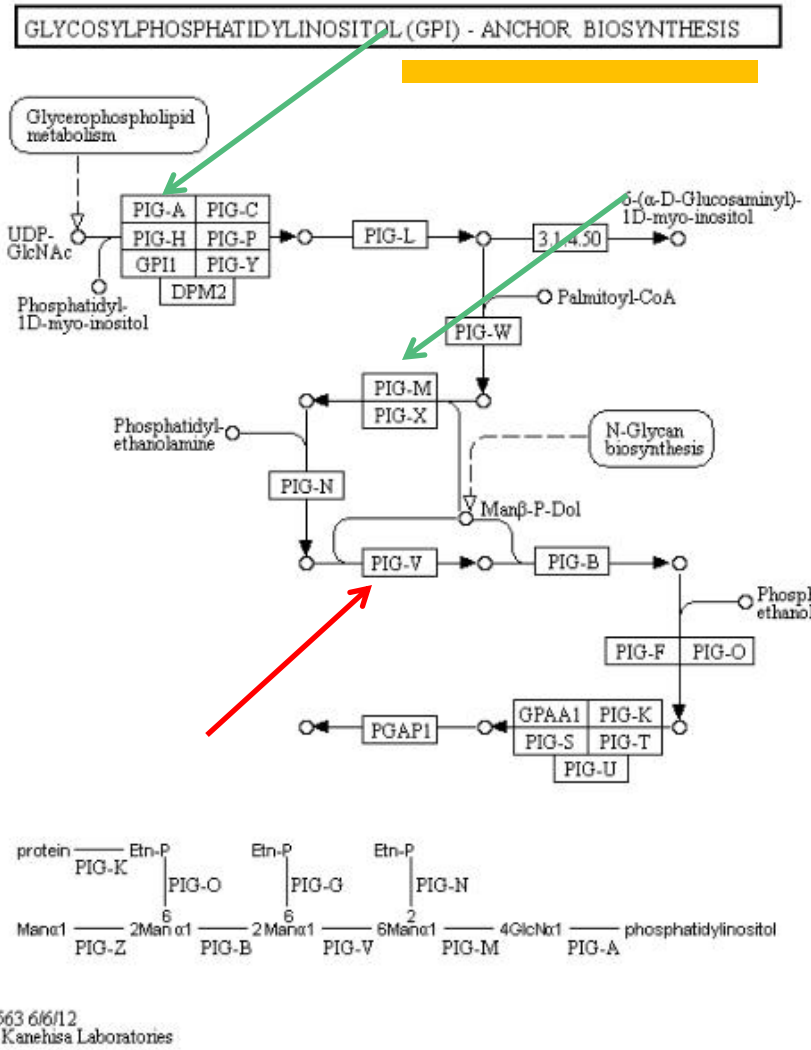
Krawitz PM, et al. Nat Genet(2010)

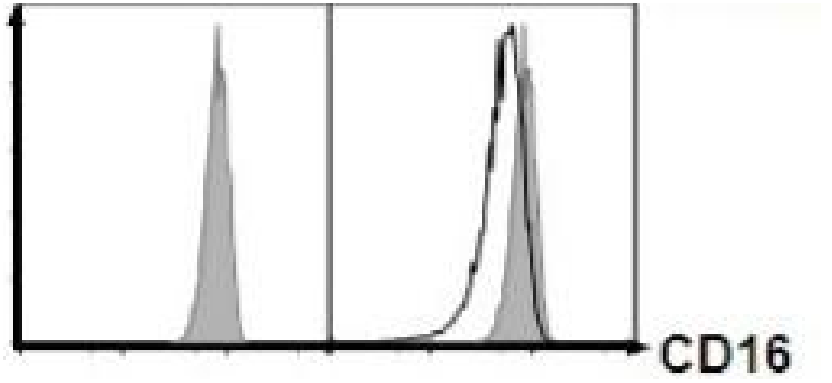
IBD application in autosomal recessive diseases

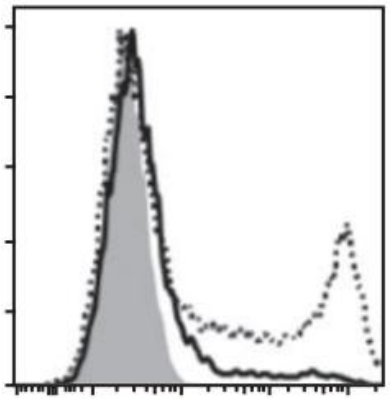
- Three siblings of nonconsanguineous parents
- ABI SOLiD platform, and Agilent's SureSelect whole-exome enrichment
- Variants shared by three affected persons, not in dbSNP130 and HapMap, 14 variants left
- Algorithm based on Hidden Markov Model decreased the search space, reducing the number of candidate genes to 2
- Mutations of PIGV, one of the candidate genes, were detected in affected persons from 3 other unrelated families

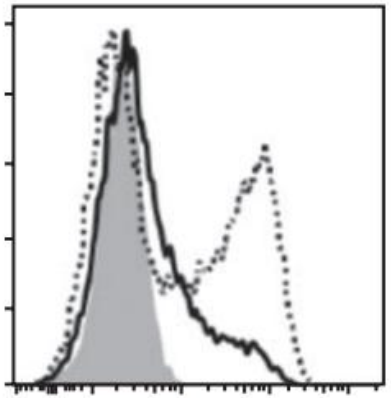
Family	cDNA	Chromosome	Protein
A	c.[1022C>A]+[1022C>A]	chr1:26994134C>A	p.[A341E]+[A341E]
B	c.[1022C>A]+[1154C>A]	chr1:26994134C>A,chr1:26994266C>A	p.[A341E]+[H385P]
C	c.[766C>A]+[766C>A]	chr1:26993878C>A	p.[Q256K]+[Q256K]
D	c.[1022C>A]+[1022C>T]	chr1:26994134C>A,chr1:26994134C>T	p.[A341E]+[A341V]

IBD application in autosomal recessive diseases









IBD application in autosomal recessive diseases

We developed an algorithm based on a Hidden Markov Model (HMM), a type of Bayesian network that is used to infer a sequence of hidden (that is, unobservable) states. We used the HMM algorithm to identify chromosomal regions with IBD = 2 in the presence of noisy (that is, potentially erroneous) sequence data. It is not possible to measure the IBD = 2 state directly; it is only possible to determine whether the genotypes of the siblings are compatible with identity-by-state status, that is, whether each sibling has the same homozygous or heterozygous genotype, a situation which we refer to as IBS*. In our model, every genetic locus was either IBD = 2 or IBD ≠ 2. The HMM was then used to predict the most likely sequence of IBD = 2 or IBD ≠ 2 chromosomal segments on the basis of the observed exome sequences of two or more affected siblings (**Supplementary Fig. 1** and **Supplementary Methods**).

Introduction to Hidden Markov model

State: Fair Biased

Symbol: $\frac{1}{2}$ Head $\frac{1}{2}$ Tail $\frac{3}{4}$ Head $\frac{1}{4}$ Tail

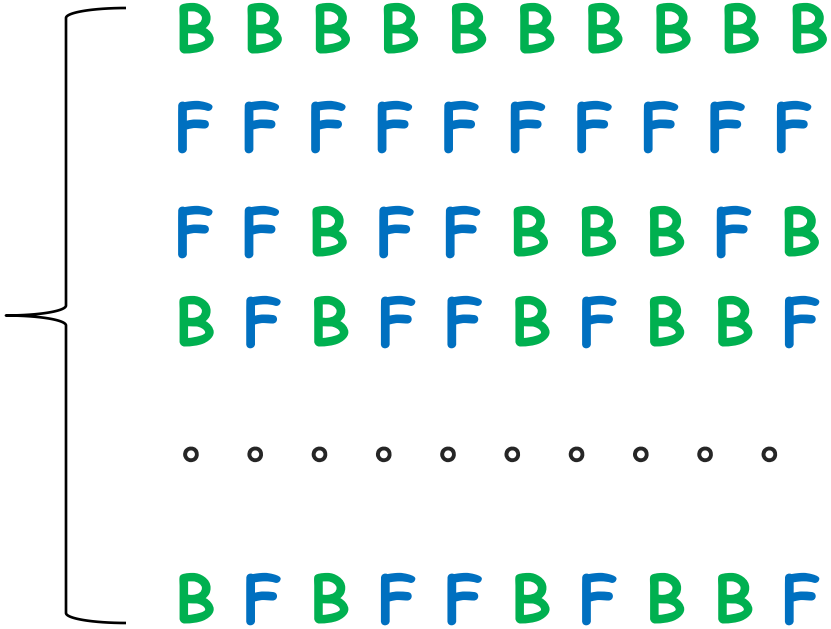
Introduction to Hidden Markov model

blackbox

H T T H T H H H T

2^10 possibility of hidden state

Which is the most likely hidden state?



Introduction to Hidden Markov model

∑: an alphabet of emitted symbols — Head and Tail

States: a set of hidden states — Fair and Biased

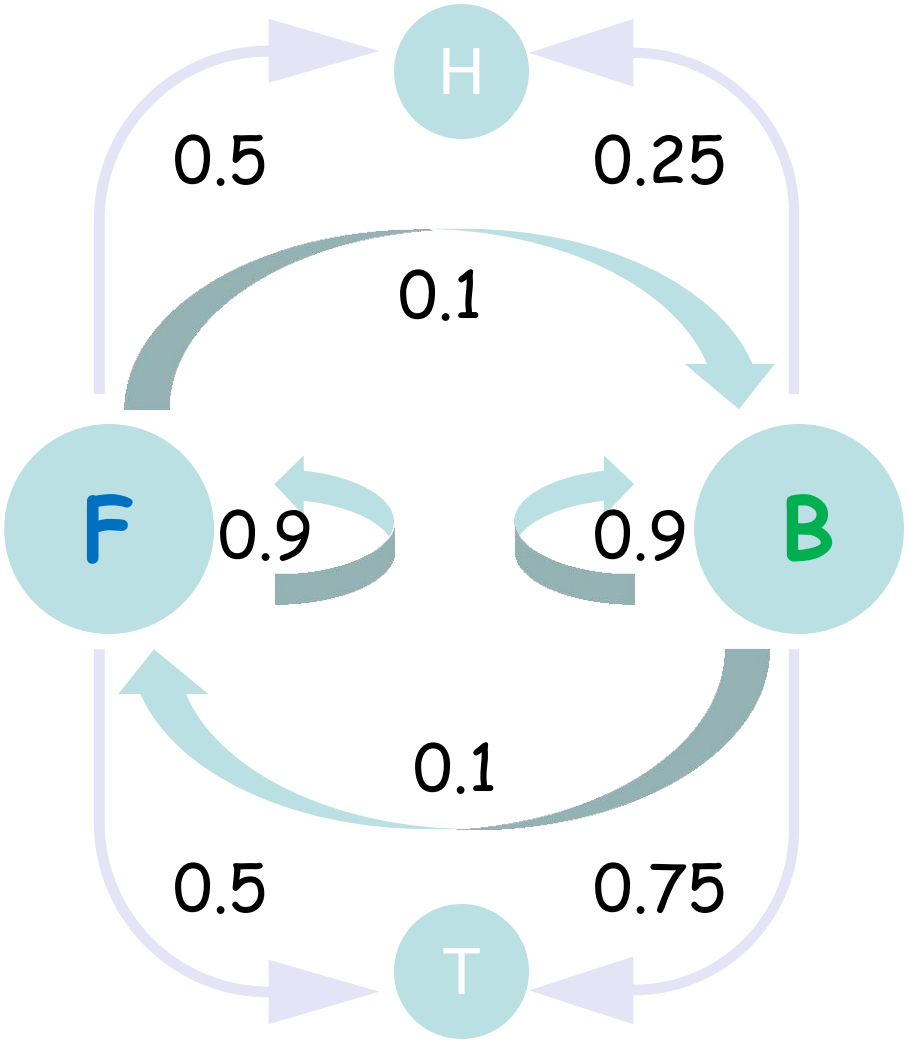
Transition = (transition$_{i,k}$)
|State| x |State|
Matrix of transition probabilities

	F	B
F	0.9	0.1
B	0.1	0.9

Emission = (emission$_k$(b))
|State| x | ∑ |
Matrix of emission probabilities

	H	T
F	0.5	0.5
B	0.75	0.25

Introduction to Hidden Markov model



Transition:

	F	B
F	0.9	0.1
B	0.1	0.9

Emssion:

	H	T
F	0.5	0.5
B	0.75	0.25

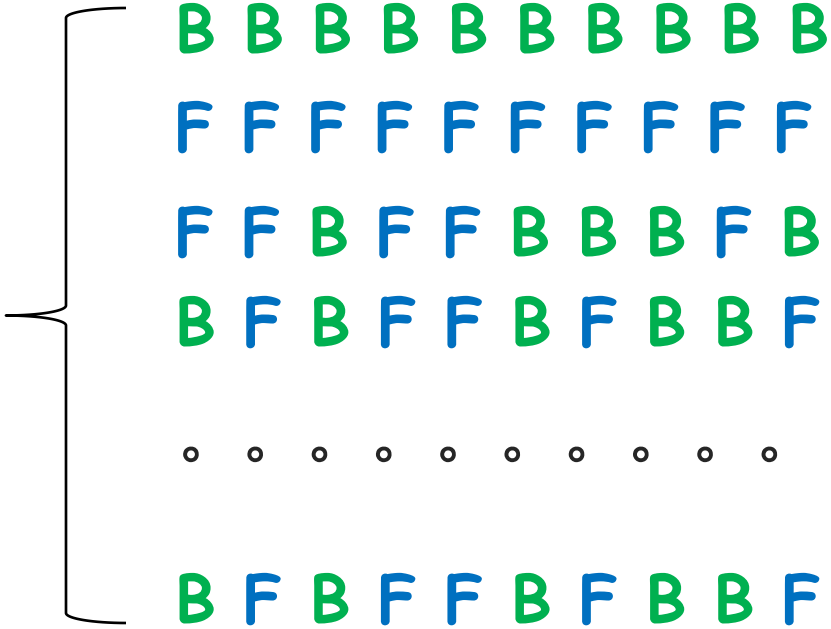
Introduction to Hidden Markov model

blackbox

H T T H T H H H T

2^10 possibility of hidden state

Which is the most likely hidden state?



Introduction to Hidden Markov model

- Hidden Path: the sequence $\pi = \pi_1 \pi_2 \ldots \pi_n$ of states that HMM passes through.
- $Pr(x, \pi)$: the probability that an HMM follows the hidden path π and emits the string $x = x_1 x_2 \ldots x_n$

π : F F B F F B B B B F

x : T T H H T H H H H T

- $Pr(x|\pi)$: the conditional probability that an HMM emits the string x after follows the hidden path π.
- $Pr(x, \pi)$是x和π的联合概率，$Pr(x|\pi)$是条件概率
- $Pr(x, \pi) = Pr(x|\pi) * Pr(\pi)$

Introduction to Hidden Markov model

$$\Pr(x, \pi) = \Pr(x|\pi) * \Pr(\pi)$$

$\Pr(x_i|\pi_i)$: probability that x_i was emitted from the state π_i (equal to emission $\pi_i(x_i)$).

	1	2	3	4	5	6	7
π:	F	F	B	F	F	B	B
x:	T	T	H	H	T	T	H
$\Pr(\pi_i \to \pi_{i+1})$:	0.5	0.9	0.1	0.1	0.9	0.1	0.9
$\Pr(x_i|\pi_i)$:	0.5	0.5	0.75	0.5	0.5	0.25	0.75

Introduction to Hidden Markov model

$$\Pr(x, \pi) = \Pr(x|\pi) * \Pr(\pi)$$

$\Pr(x_i|\pi_i)$: probability that x_i was emitted from the state π_i (equal to emission $\pi_i(x_i)$).

	1	2	3	4	5	6	7
π:	F	F	B	F	F	B	B
x:	T	T	H	H	T	T	H

$\Pr(\pi_i \to \pi_{i+1})$: 0.5 * 0.9 * 0.1 * 0.1 * 0.9 * 0.1 * 0.9

$\Pr(x_i|\pi_i)$: 0.5 * 0.5 * 0.75 * 0.5 * 0.5 * 0.25 * 0.75

$$\Pr(\pi) = \prod_{i=1,n} \Pr(\pi_i \to \pi_{i+1}) = \prod_{i=1,n} \text{transition}\ \pi_i \to \pi_{i+1}$$

$$\Pr(x|\pi) = \prod_{i=1,n} \Pr(x_i|\pi_i) = \prod_{i=1,n} \text{emission}\ \pi_i(x_i)$$

Introduction to Hidden Markov model

$$\Pr(x, \pi) = \Pr(x|\pi) * \Pr(\pi)$$

$\Pr(x_i|\pi_i)$: probability that x_i was emitted from the state π_i (equal to emission $\pi_i(x_i)$).

	1	2	3	4	5	6	7
π :	F	F	B	F	F	B	B
x:	T	T	H	H	T	T	H

$\Pr(\pi_i \to \pi_{i+1})$:	0.5	*	0.9	*	0.1	*	0.1	*	0.9	*	0.1	*	0.9
$\Pr(x_i\|\pi_i)$:	0.5	*	0.5	*	0.75	*	0.5	*	0.5	*	0.25	*	0.75

$$\Pr(\pi) = \prod_{i=1,n} \Pr(\pi_i \to \pi_{i+1}) = \prod_{i=1,n} \text{transition}\ \pi_i \to \pi_{i+1}$$

$$\Pr(x|\pi) = \prod_{i=1,n} \Pr(x_i|\pi_i) = \prod_{i=1,n} \text{emission}\ \pi_i(x_i)$$

Introduction to Hidden Markov model

- Decoding Problem: Find an optimal hidden path in an HMM given its emitted string.
- Input: A string $x = x_1 x_2 \ldots x_n$ emitted by an HMM (Σ, State, Transition, Emission)
- Output: A path π that maximizes the probability $Pr(x, \pi)$ over all possible paths through this HMM.

$$Pr(x, \pi) = Pr(x|\pi) * Pr(\pi)$$
$$= \prod_{i=1,n} Pr(x_i|\pi_i) * Pr(\pi_i \rightarrow \pi_{i+1})$$
$$= \prod_{i=1,n} \text{emission}_{\pi_i}(x_i) * \text{transition}_{\pi_i \rightarrow \pi_{i+1}}$$

- The process of discovering the sequence of hidden states, given the sequence of observations, is known as decoding. The **Viterbi** algorithm is commonly used for decoding.

Hidden Markov model and IBD

∑: an alphabet of emitted symbols — Head and Tail

States: a set of hidden states — Fair and Biased

Transition = $(\text{transition}_{i,k})$
|State| x |State|
Matrix of transition probabilities

	F	B
F	0.9	0.1
B	0.1	0.9

Emission = $(\text{emission}_{k}(b))$
|State| x | ∑ |
Matrix of emission probabilities

	H	T
F	0.5	0.5
B	0.75	0.25

Hidden Markov model and IBD

Σ: an alphabet of emitted symbols

States: a set of hidden states

Transition = $(\text{transition}_{i,k})$
|State| x |State|
Matrix of transition probabilities

Emission = $(\text{emission}_k(b))$
|State| x | Σ |
Matrix of emission probabilities

$\Pr(x, \pi)$?

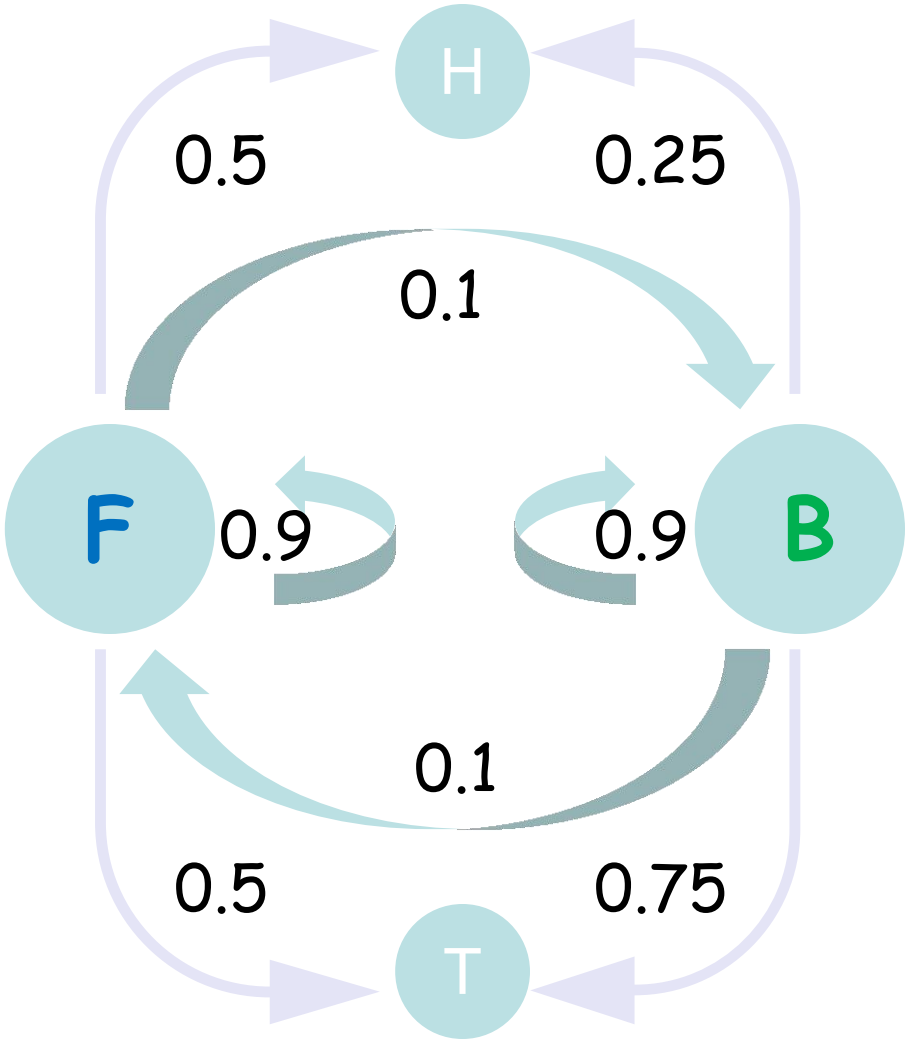
IBS = 2 and IBS ≠ 2

IBD = 2 and IBD ≠ 2

?

?

Hidden Markov model and IBD



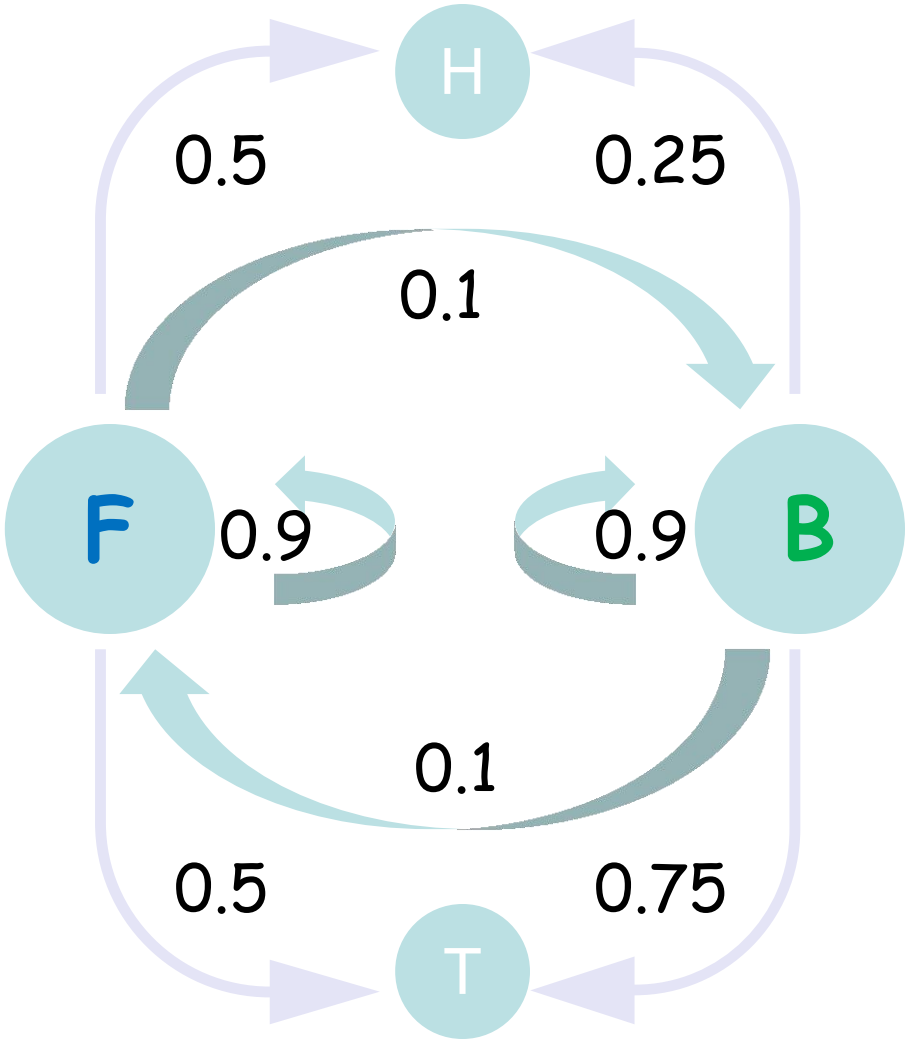
Transition:

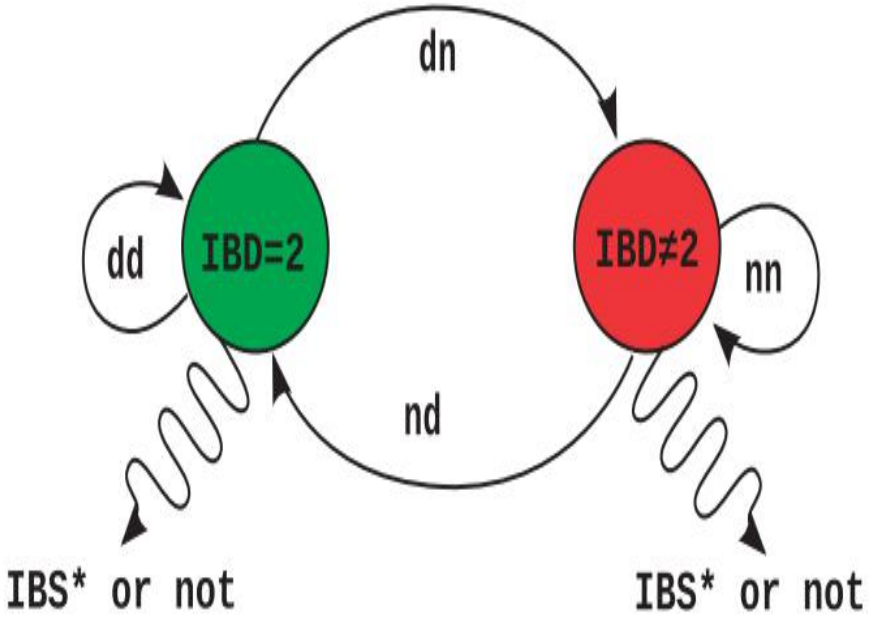
	F	B
F	0.9	0.1
B	0.1	0.9

Emssion:

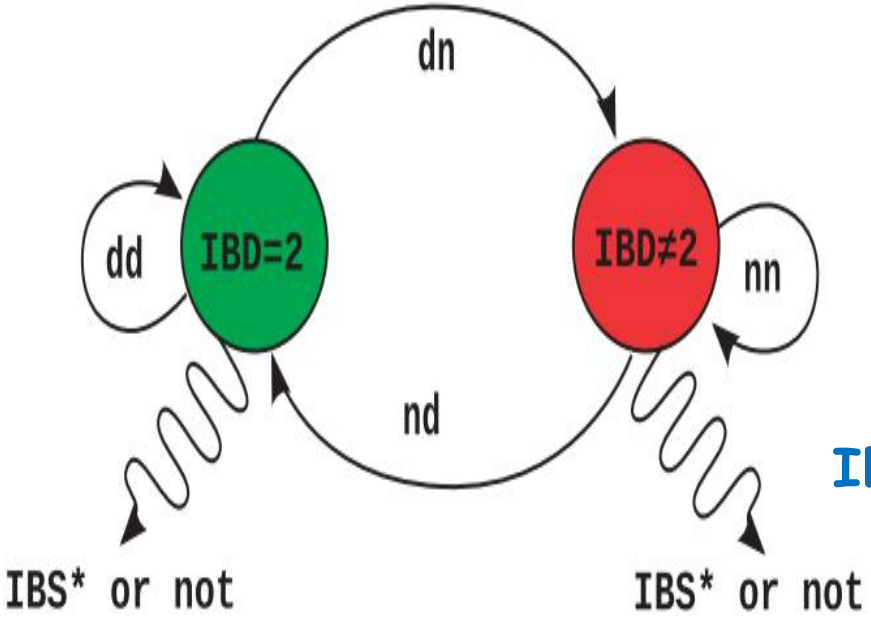
	H	T
F	0.5	0.5
B	0.75	0.25

Hidden Markov model and IBD





Hidden Markov model and IBD



Transition:

$$T(x_i, x_j, s) = \begin{bmatrix} nn & nd \\ dn & dd \end{bmatrix}$$

Emission:

IBD = 2

$$e_D(1) = (1-\varepsilon)^{2n}$$

$$e_D(0) = 1-(1-\varepsilon)^{2n}$$

IBD ≠ 2

$e_N(1) = 0.36$

$e_N(0) = 1 - 0.36$ for n = 2

$e_N(1) = 0.26$

$e_N(0) = 1 - 0.26$ for n = 3

Hidden Markov model and IBD

The input to the model is the sequence of l observed sequence variants x, the set of the hidden states h including the initial state, a matrix of emission probabilities e, and a transition rate matrix T. A path π defines a sequence of IBD=2/non-IBD=2 states that could have generated the observations. The joint probability of a path π and a sequence of L observations is given by:

$$P(x,\pi) = t_{0\pi_1} \prod_{L}^{i=1} e_{\pi_i}(x_i) t_{\pi_{i-1}\pi_i},$$

where are the respective transition probabilities and the probability states are initialized to the *a priori* probability of being in states D and N from equation (1). The most probable path π^* is calculated separately for each autosomal chromosome as:

$$\pi^* = \arg\max_{\pi} P(x,\pi)$$

π^* can be found recursively using Viterbi's algorithm, adapted to a inhomogeneous Markov model, which takes into account that the transition rates $t_{\pi_{i-1}\pi_i}$ not only depend on π_{i-1} and π_i but also on the recombination rate between the observation x_i and x_{i-1} . The states of the most probable path π^* indicate the predicted IBD=2 and non-IBD=2 chromosomal segments. Intersection filters can now be applied to the genes located in the IBD=2 regions to search for the disease gene.

Thanks

Jiang Chongyi