

CARLOS HENRIQUE MADEIROS CASTELLETTI

**REGIÕES GÊNICAS DE INTERESSE PARA O DESENVOLVIMENTO
DE UM DIAGNÓSTICO DO PAPILLOMAVIRUS HUMANO:
*ABORDAGEM IN SILICO***

RECIFE

2006

CARLOS HENRIQUE MADEIROS CASTELLETTI

**REGIÕES GÊNICAS DE INTERESSE PARA O DESENVOLVIMENTO
DE UM DIAGNÓSTICO DO PAPILLOMAVIRUS HUMANO:
ABORDAGEM *IN SILICO***

Dissertação apresentada ao Programa de Pós-Graduação do Centro de Ciências Biológicas da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciências Biológicas, ênfase em Biotecnologia.

Orientador: José Luiz de Lima Filho

RECIFE

2006

Castelletti, Carlos Henrique Madeiros

Regiões gênicas de interesse para o desenvolvimento de um diagnóstico do Papillomavirus Humano: abordagem *in silico*/Carlos Henrique Madeiros Castelletti. – Recife: O Autor, 2006.

71 folhas: il.

Dissertação (Mestrado em Ciências Biológicas) – UFPE. CCB, 2006.

1.Papillomavirus 2. Câncer 3. Câncer de Colo de Útero I.Título

616-006

CDU (2.ed.)

UFPE

616.994

CDD (22.ed.)

CCB – 2007 - 121

CARLOS HENRIQUE MADEIROS CASTELLETTI

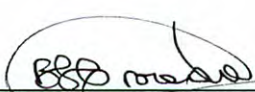
**REGIÕES GÊNICAS DE INTERESSE PARA O DESENVOLVIMENTO
DE UM DIAGNÓSTICO DO PAPILLOMAVIRUS HUMANO:
ABORDAGEM *IN SILICO***

Aprovado em 26 de junho de 2006, pela comissão examinadora.

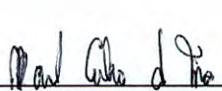
COMISSÃO EXAMINADORA



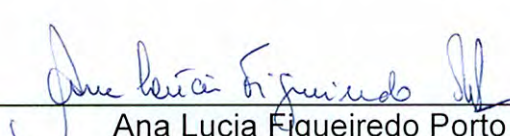
Prof. Dr. José Luiz de Lima Filho (Orientador)
Departamento de Bioquímica - UFPE



Benildo Sousa Cavada
Departamento de Bioquímica e Biologia Molecular - UFC



Manoel Eusébio de Lima
Centro de Informática - UFPE



Ana Lucia Figueiredo Porto (Suplente)
Departamento de Morfologia e Fisiologia Animal - UFRPE



Gandhi Rádis Baptista (Suplente)
Departamento de Bioquímica - UFPE

RECIFE

2006

Dedico esta dissertação a Vera e Carlos (meus pais), Cristina e Eduardo (meus irmãos) e Bianca e Pietro (meus sobrinhos).

AGRADECIMENTOS

Aos meus pais, Vera e Carlos, por todo o apoio, compreensão, cuidado e carinho, não somente nesta etapa, mas durante todo o percurso até aqui. A Cristina e Eduardo, meus irmãos, por sempre estarem por perto. A Bianca e Pietro, meus sobrinhos, que me divertiram nos momentos de descanso.

À Maíra Mafra, minha namorada, pela extraordinária paciência, compreensão, companheirismo e incentivo nos momentos intranquilos e de apreensão. Pelo olhar carinhoso que levo sempre comigo nos momentos em que não estamos juntos.

Ao Prof. Dr. Jose Luiz de Lima Filho, pelo incentivo, oportunidade e confiança depositada quando me chamou para desenvolvermos juntos, um novo grupo de pesquisas dentro do Lika.

Ao David Moraes, que acreditou na idéia deste trabalho e não mediu esforços em ajudar a concretizá-lo juntos. Brigamos, discutimos, conversamos, e fizemos.

À Danyelly Bruneska, pela ajuda e incentivo constante e pelas contribuições relevantes a este trabalho.

Aos membros das bancas - de qualificação e final - pelas relevantes contribuições para este trabalho.

Ao Lika, ao qual retorno, por disponibilizar sua infra-estrutura e equipamentos para a realização deste trabalho. A seus funcionários aos quais recorri diversas vezes.

Ao Laboratório de Bioinformática do LNCC, em especial a Profa. Dra. Ana Tereza Vasconcelos pela disponibilidade dos computadores de grande porte e ao Prof. Dr. Raul Fonseca Neto que desenvolveu a parte matemática deste trabalho.

Aos amigos, sempre curiosos em saber como estava o trabalho e que também ajudavam a “esquecer um pouco” dele.

A todos, que direta ou indiretamente contribuíram para que cumprisse mais essa etapa.

“Não tema avançar lentamente receie apenas ficar parado”.

Provérbio Chinês

“Escolha um trabalho que você ame e não terá que trabalhar um único dia da sua vida”.

Confúcio (551 A.C. – 449 A.C.)

“A coisa mais bela que o homem pode experimentar é o mistério. É essa emoção fundamental que está na raiz de toda ciência e de toda arte”.

Albert Einstein (1879 – 1955)

RESUMO

O câncer do colo do útero é o segundo câncer em número de casos entre as mulheres e o sétimo de maior incidência no mundo com 493 mil novos casos por ano, com uma taxa de mortalidade de 50 a 55%. A estimativa para o Brasil é de 19.260 novos casos para 2006, com um risco estimado de 20,31 casos a cada 100 mil mulheres. No Brasil, assim como nos outros países em desenvolvimento, a frequência do teste de Papanicolau é baixa e os casos são normalmente diagnosticados em fase avançada. Em todo o mundo o Papillomavirus Humano (HPV), diretamente associado ao câncer do colo do útero, é um problema de saúde pública. A análise do genoma, através da genômica comparativa, dos HPVs de alto e baixo risco pode indicar regiões de grande importância associadas à sua patogenicidade. O presente trabalho propõe uma metodologia capaz de classificar o risco à carcinogenicidade utilizando as seqüências de cinco proteínas, E6, E7, E1, L1 e L2 que estão presente em 78 genomas de HPV disponíveis nos bancos de dados públicos. Utilizando-se dos Modelos Ocultos de Markov (HMM) foram selecionadas 24 regiões nas cinco proteínas. Destas 24 regiões, oito foram escolhidas após aplicação de modelo discriminativo baseado em Máquinas de Vetores de Suporte (SVM) acoplado a um algoritmo genético. Diferentes combinações destas oito regiões produziram um conjunto de classificadores capazes de predizer com 100% de eficiência o tipo de risco do HPV. Esta técnica permitiu determinar regiões de potencial importância na carcinogenicidade do HPV. Tais resultados sugerem regiões gênicas de interesse para o desenvolvimento de um método de diagnóstico da presença do DNA viral, contribuindo na prevenção do câncer do colo do útero.

Palavras chaves: Papillomavirus, Diagnóstico Molecular, Bioinformática, *in silico*.

ABSTRACT

Cervical cancer is the second most common cancer in women and the seventh in the world with 493,000 of new cases per year, and mortality between 50 and 55%. In Brazil, the estimate is 19,260 new cases for 2006, with an estimated risk for 20.31 cases to each 100,000 women. In Brazil, as well as in the others developing countries, the frequency of the Pap smear are low and the cases normally are diagnosed in advanced phase. In the whole world the Human Papillomavirus (HPV), associated with cervical cancer, is a public health problem. The genome analysis, through comparative genomics, of the high and low risk HPVs, can indicate important regions associated in carcinogenicity. This study proposes a methodology to classify the carcinogenicity risk of HPVs by using protein sequences of five genes, E6, E1, E7, L1, and L2, which are present in 78 HPV genomes available in the public databases. Using Hidden Markov Models (HMM) twenty-four regions had been selected in the five proteins. Eight, of 24 regions, were selected when used a discriminative model based on Support Vector Machine (SVM) coupled with a Genetic Algorithm (GA). Different combinations of these eight regions produced a set of classifiers capable of correctly predicting 100% of HPV risk type. This technique allowed determining regions that can have an important role in the carcinogenicity of the HPV. These results suggest regions of interest for developing a method of diagnosis for viral DNA presence, contributing in the prevention of cervical cancer.

Keywords: Papillomavirus, Molecular Diagnostics, Bioinformatics, *in silico*.

LISTA DE FIGURAS

- Figura 1. Incidência (A) e mortalidade (B) por câncer do colo do útero ajustado por idade para 100.000 mulheres. Dados baseados no Globocan 2002, IARC (<http://www-dep.iarc.fr>).19
- Figura 2. Taxa do número médio de anos potenciais perdidos por câncer do colo do útero (Inca, 2002).21
- Figura 3. Taxa de incidência por grupo de idade para 100.000 mulheres para o Brasil e a Inglaterra. Adaptado de (Bosch e de Sanjose, 2003).24
- Figura 4. A ilustração mostra uma vista da superfície molecular do modelo atômico do papillomavirus proposta por (Modis *et al.*, 2002).29
- Figura 5. Ilustrações de dois tipos de Papillomavirus humano. A - Organização simplificada do genoma do HPV-16 (Linearizado). B - Organização circular do genoma do HPV 11. Os E genes codificam as proteínas não estruturais. Os L genes codificam as duas proteínas do capsídio viral.30
- Figura 6. Patogenicidade do HPV. Os genes E6 e E7 codificam proteínas multifuncionais que se ligam primariamente às proteínas p53 e pRb, alterando o ciclo celular. Modificado de (Burd, 2003).32
- Figura 7. Árvore filogenética montada a partir da seqüência do gene L1 de 118 tipos de Papillomavirus humano (de Villiers *et al.*, 2004).35
- Figura 8. Prevalência cumulativa dos tipos mais comuns de HPVs associados a casos de câncer cervical. Adaptado de (Bosch e de Sanjose, 2003).38
- Figura 9. Representação do hiperplano ótimo, margem maximizada, do SVM. Dentro do cubo representantes das duas classes (vermelho e verde) separados nas margens opostas.43

LISTA DE TABELAS

Tabela 1. Incidência e mortalidade dos principais tipos de câncer em mulheres em todo mundo, Globocan 2002.	16
Tabela 2. Estimativa de sobrevivência em porcentagem para 6 tipos de câncer por país/área (ajustado por idade).....	17
Tabela 3. Número de casos, porcentagem e taxa bruta para o câncer do colo do útero.	21
Tabela 4. Classificação epidemiológica dos tipos de Papillomavirus Humano associada ao alto e baixo risco para o câncer do colo do útero.....	37

LISTA DE ABREVIATURAS E SIGLAS

AG – Algoritmo Genético

CIN – Neoplasias Cervicais

DNA – Ácido Desoxirribonucléico

Globocan – Global Cancer Database

HIV – Vírus da Imunodeficiência Humana

HPV – Papillomavirus Humano

HMM – Modelos Ocultos de Markov

HSV2 – Vírus da Herpes Simples 2

IARC – Agência Internacional para Pesquisa em Câncer

Inca – Instituto Nacional de Câncer

IST – Infecções Sexualmente Transmissíveis

LCR – Região Longa de Controle

ML – Aprendizado de Máquina

OMS – Organização Mundial da Saúde

ORF – Código Aberto de Leitura

PB – Pares de Base

PCR – Reação em Cadeia de Polimerase

SVM – Máquinas de Vetores de Suporte

UVB – Raios Ultravioleta B

SUMÁRIO

Introdução	14
Justificativa e Objetivos	15
Revisão da literatura	16
1. O câncer no mundo – Estatísticas globais	16
1.1. Câncer do colo do útero – Redução nos casos, não na mortalidade	17
2. Estatísticas Brasileiras, Nordestinas, Pernambucanas	20
2.1. O câncer do colo do útero no Brasil	20
3. Câncer do colo do útero	22
3.1. Fatores de Risco	22
3.2. Epidemiologia	23
3.3. Diagnóstico	25
3.4. Prevenção e tratamento	27
4. Papillomavirus humano (HPV)	28
4.1. Estrutura	28
4.2. Patogenicidade	30
4.3. Classificação	33
4.3.1. Filogenia	34

4.3.2. Alto e baixo risco associado ao câncer cervical	36
4.4. Os tipos mais comuns associados à oncogenicidade	37
5. Abordagem <i>in silico</i>	39
5.1. Bioinformática – Ciência multidisciplinar na busca de informações	39
5.2. Modelos Ocultos de Markov – HMM	41
5.3. Aprendizado de Máquina – ML	42
Referências Bibliográficas	45
Artigo - Inference of Carcinogenicity of Human Papillomaviruses Based on Hidden Markov Model and Support Vector Machines.....	57
Conclusões e Perspectivas	67
Anexo – Atividades relevantes desenvolvidas durante o mestrado.....	69

INTRODUÇÃO

O câncer é uma das maiores causa de morte no mundo e o Brasil acompanha a tendência mundial para novos casos de câncer. Segundo o Instituto Nacional de Câncer, Inca, as estimativas para o ano de 2006 apontam para quase meio milhão de novos casos.

O câncer do colo do útero ou câncer cervical é o segundo em número de casos entre as mulheres e o sétimo de maior incidência no mundo, com 493 mil novos casos por ano e mortalidade de 50%. No Brasil a estimativa é de 19.260 novos casos com um risco estimado de 20,31 casos a cada 100 mil mulheres.

Estudos epidemiológicos associaram a progressão de lesões de baixo risco ao câncer do colo do útero a vários fatores de risco. O principal fator associado é a detecção da presença do Papillomavirus Humano (HPV).

A prevenção é uma das melhores formas para o controle das Infecções sexualmente transmissíveis como é o caso do HPV. Desta forma as mesmas atitudes tomadas em relação a outras doenças podem ajudar no combate à transmissão e infecção pelo HPV.

O teste preventivo de Papanicolau continua sendo o principal método de prevenção do câncer do colo do útero. Quando diagnosticadas displasias brandas ou moderadas, em estágios menos avançados, apresentam alta taxa de regressão nos primeiros dois anos de diagnóstico.

JUSTIFICATIVA E OBJETIVOS

Em todo o mundo o Papillomavirus humano, associado ao câncer do colo do útero, é um problema de saúde pública. No Brasil, assim como nos outros países em desenvolvimento, a frequência do teste de Papanicolau é baixa e os casos são normalmente diagnosticados em fase avançada.

As técnicas de biologia molecular – aliadas as novas abordagens computacionais –, *in silico*, são capazes de produzir diagnósticos mais precisos em busca de uma melhor qualidade de vida da população, combatendo este câncer que mata em 50% dos casos.

A análise do genoma, através da genômica comparativa, dos HPVs de alto e baixo risco pode indicar regiões de grande importância associadas à sua patogenicidade separando-os em duas classes distintas.

Diante do fato de que é possível diminuir a incidência do câncer do colo do útero com prevenção através do uso de metodologias já consagradas, auxiliadas agora por técnicas moleculares de detecção de DNA, e ainda, utilizando-se as novas abordagens *in silico* para diminuir o tempo e o custo de novas formas de diagnósticos propomos como objetivo deste trabalho:

Desenvolver uma nova metodologia capaz de classificar o risco à carcinogenicidade dos tipos de HPV utilizando as seqüências de cinco proteínas, E6, E7, E1, L1 e L2 que estão presente em 78 genomas de HPV disponíveis nos bancos de dados públicos.

Este projeto dará suporte de novos conhecimentos a projetos de biologia molecular que serão desenvolvidos posteriormente.

REVISÃO DA LITERATURA

1. O CÂNCER NO MUNDO – ESTATÍSTICAS GLOBAIS

Com base nos dados do GLOBOCAN 2002, da Agência Internacional para Pesquisa em Câncer (IARC), foram estimados 10,9 milhões de novos casos de câncer em todo o mundo para 2006. Estima-se ainda 6,7 milhões de mortes e 24,6 milhões de pessoas convivendo com o câncer (até 5 anos após o diagnóstico). Dependendo do foco de interesse, incidência ou mortalidade, os tipos de câncer podem ocupar posições diferentes na tabela (Tabela 1) (Jemal *et al.*, 2006; Parkin *et al.*, 2005).

A incidência e a mortalidade variam de acordo com o país/área estudado e entre os sexos. Em países considerados desenvolvidos a taxa de sobrevivência é, em média, maior que a taxa de países em desenvolvimento. Entre os sexos, os homens têm melhor índice de sobrevivência nos países desenvolvidos que nos países em desenvolvimento, ao contrário das mulheres que têm melhor taxa de sobrevivência nos países em desenvolvimento (Tabela 2) (Parkin *et al.*, 2005).

Tabela 1. Incidência e mortalidade dos principais tipos de câncer em mulheres em todo mundo, Globocan 2002.

	Incidência		Mortalidade		Taxa de fatalidade	
	Mulheres	Total	Mulheres	Total	Mulheres	Total
Estômago	330.518	933.937	254.297	700.349	0,77	0,75
Cólon/Reto	472.687	1.023.152	250.532	528.978	0,53	0,52
Pulmão	386.891	1.352.132	330.786	1.178.918	0,85	0,87
Mama	1.151.298	1.151.298	410.712	410.712	0,36	0,36
Colo do útero	493.243	493.243	273.505	273.505	0,55	0,55
Corpo uterino	198.783	198.783	50.327	50.327	0,25	0,25
Ovário	204.499	204.499	124.860	124.860	0,61	0,61
Todos *	5.060.657	10.862.496	2.927.896	6.723.887	0,58	0,62

* exceto câncer de pele. Modificado de (Parkin *et al.*, 2005)

Tabela 2. Estimativa de sobrevivência em porcentagem para 6 tipos de câncer por país/área (ajustado por idade).

	Áreas desenvolvidas					Am Sul	IND	THA	Áfr Sub	Todas* áreas
	USA	EU Ori	EU Oci	JPN	Todas* áreas					
Estômago (H)	44	15	30	54	35	25	14	12	7	21
Estômago (M)	33	16	24	51	31	24	14	14	5	20
Cólon/Reto (H)	66	35	56	65	56	50	28	37	13	39
Cólon/Reto (M)	65	36	53	58	54	50	31	37	14	39
Pulmão (H)	21	9	9	15	13	8	12	5	4	12
Pulmão (M)	26	10	14	22	20	1	11	5	5	12
Mama	81	58	74	75	73	67	46	62	32	57
Colo do útero	70	51	66	65	61	55	42	58	21	41
Corpo uterino	89	69	83	79	82	70	59	67	61	67

* Todas as áreas corresponde à média das áreas estudadas, das quais destacamos apenas oito. USA – Estados Unidos, EU Ori – Europa Oriental, EU Oci – Europa Ocidental, JPN – Japão, Am Sul – América do Sul, IND – Índia, THA – Tailândia, Afr Sub – África Subsaariana. Adaptado de (Parkin *et al.*, 2005)

1.1. Câncer do colo do útero – Redução nos casos, não na mortalidade

O câncer do colo do útero ou câncer cervical é o segundo em número de casos entre as mulheres e o sétimo de maior incidência no mundo com 493 mil novos casos por ano, sendo sua taxa de fatalidade de 50 a 55% (Jemal *et al.*, 2006; Inca, 2005; Parkin *et al.*, 2005).

Possui uma taxa de sobrevivência, em média, de 61% em países desenvolvidos e 41% nos países em desenvolvimento, uma vez que nestes países os casos são diagnosticados relativamente em estágio avançado. É responsável por 15% dos cânceres em mulheres nos países em desenvolvimento, sendo sua maior incidência na África Subsaariana, Melanésia, América Latina e Caribe, Ásia Central e Sudeste da Ásia. Estas regiões correspondem a 83% de todos os casos no mundo. Nos países desenvolvidos a ocorrência é de apenas 3,8% do total de cânceres femininos (Figura 1) (Jemal *et al.*, 2006).

Na década de 90 do século XX, o câncer cervical era o primeiro câncer entre as mulheres, quando o câncer de mama tornou-se predominante. O uso constante de testes preventivos na triagem de pacientes com alterações celulares tornou-se mais freqüente nos últimos 50 anos, o que acarretou a redução dos casos de incidência por este câncer (Franco *et al.*, 2001).

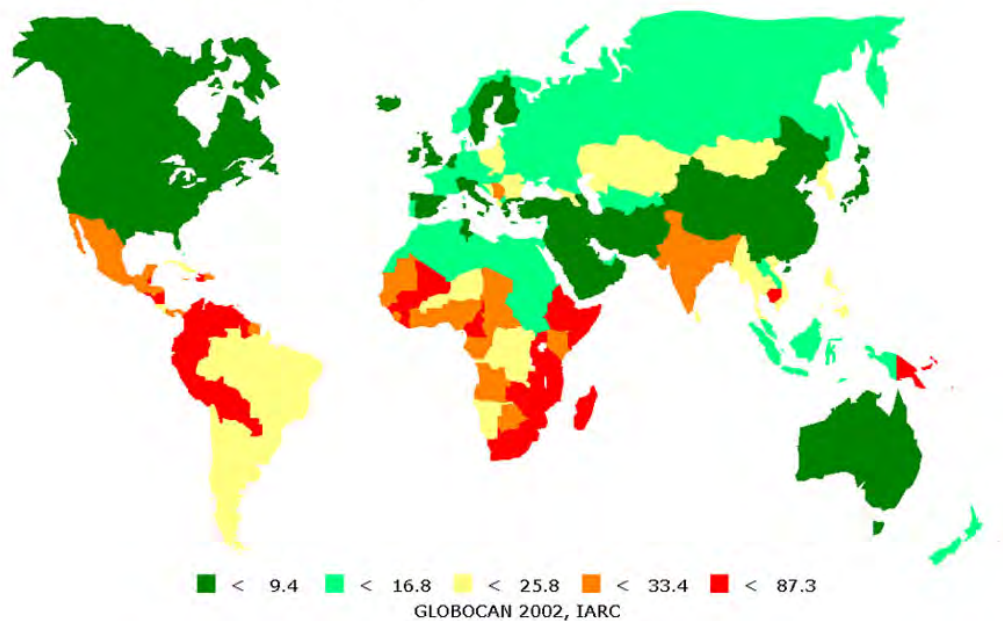
Esta redução também ocorreu por um progresso substancial – nos últimos 20 anos – do conhecimento adquirido sobre a história natural do câncer do colo do útero e avanços nos principais tratamentos. Entretanto estas mudanças são mais evidentes nos chamados países desenvolvidos devido a uma conscientização e melhora nos programas de prevenção, que se tornaram mais organizados e eficientes (Franco *et al.*, 2001).

Comparando-se os dados brutos sobre os diversos cânceres entre 2000 e 2006 da população dos Estados Unidos nota-se um aumento no número total de casos. Entretanto o número de mulheres acometidas pelo câncer do colo do útero vem gradativamente diminuindo. O mesmo não ocorre com a mortalidade pelo câncer do colo do útero que vem aumentando ano a ano. Em 2002 a taxa de mortalidade era de 31% e para 2006 a expectativa é de 38%. A relação do câncer do colo do útero em comparação a outros cânceres nas mulheres em 2000 era de 2,13% com uma perspectiva de 1,43% para o ano de 2006 (Jemal *et al.*, 2006; Jemal *et al.*, 2005; Jemal *et al.*, 2004; Jemal *et al.*, 2003; Jemal *et al.*, 2002; Greenlee *et al.*, 2001; Greenlee *et al.*, 2000).

Entretanto esta redução é em relação ao número de casos de câncer cervical invasivo e carcinoma cervical de células escamosas. Em seu estudo mais recente Castellsague (2006) demonstra que o número de mulheres jovens com adenocarcinomas cervicais cresce nos países desenvolvidos. Nos Estados Unidos a

relação entre adenocarcinomas e carcinoma de células escamosas e todos os cânceres cervicais dobrou em pouco mais de 20 anos. Este fato indica que os atuais programas de triagem podem ser insuficientes para a detecção desta parcela de lesões precursoras do câncer do colo do útero.

A - Incidência do câncer do colo do útero



B - Mortalidade do câncer do colo do útero

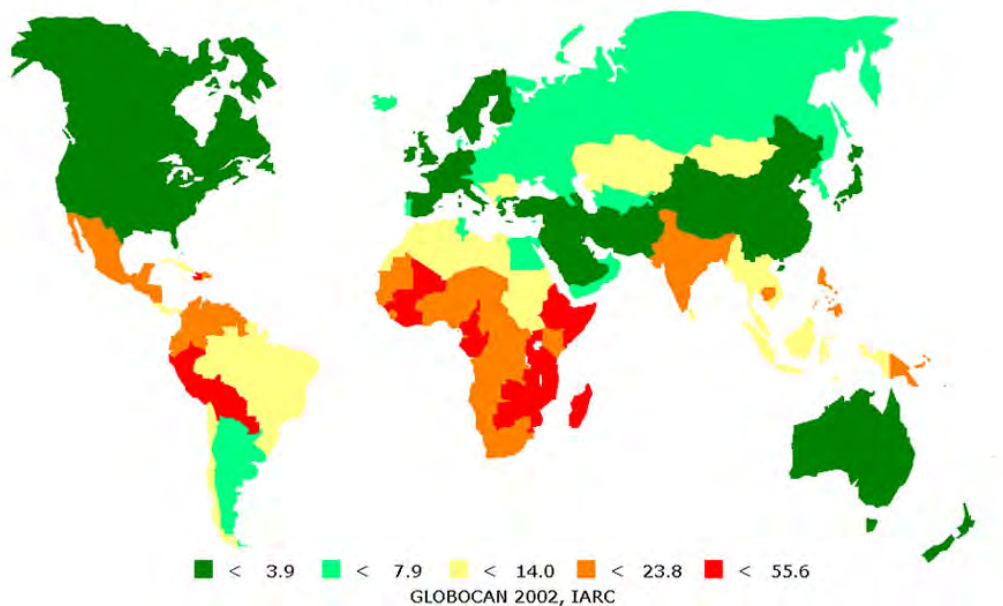


Figura 1. Incidência (A) e mortalidade (B) por câncer do colo do útero ajustado por idade para 100.000 mulheres. Dados baseados no Globocan 2002, IARC (<http://www-dep.iarc.fr>).

2. ESTATÍSTICAS BRASILEIRAS, NORDESTINAS, PERNAMBUCANAS

O Brasil acompanha a tendência mundial para novos casos de câncer e uma ordem similar na incidência de câncer é observada com e sem distinção de sexo.

2.1. O câncer do colo do útero no Brasil

A incidência do câncer do colo do útero no Brasil será de 19.260 novos casos para 2006. É o sexto de maior incidência sem distinção de sexo e o segundo entre as mulheres, com um risco estimado de 20,31 casos a cada 100 mil mulheres para todo o Brasil. Nas capitais este número aumenta para 27,11 para cada 100 mil. Estes novos casos correspondem a 10,92% dos cânceres femininos no Brasil, deixando o país em um patamar médio entre os países desenvolvidos e em desenvolvimento (Inca, 2005).

As regiões Norte e Nordeste são as mais atingidas. A região Nordeste corresponde a 22,90% do câncer do colo do útero do Brasil, com 4.410 novos casos para 2006 e uma taxa bruta de 16,75 para cada 100 mil mulheres, pouco abaixo da média nacional (Inca, 2005).

As capitais nordestinas têm uma taxa bruta superior à média nacional com 25,81 para cada 100 mil, e um total de 1.520 casos correspondendo a 7,9% dos casos brasileiros (Inca, 2005).

Pernambuco, juntamente com a Bahia, é o estado do Nordeste com maior incidência deste tipo de câncer e o sexto dos 26 estados brasileiros, com 5,04% no âmbito nacional. Sua taxa bruta é pouco superior à média nacional com 22,16 e sua capital, Recife, tem uma taxa bruta de 29,66 casos por cada 100 mil mulheres (Tabela 3) (Inca, 2005).

Tabela 3. Número de casos, porcentagem e taxa bruta para o câncer do colo do útero.

	Nº casos	% Nacional	% Regional	Taxa Bruta Total	Taxa Bruta Capital
Brasil	19.260	100		20,31	27,11
Nordeste	4.410	22,90	100	16,75	25,81
Pernambuco	970	5,04	22,00	22,16	29,66

Modificado de (Inca, 2005). Taxa Bruta por 100 mil mulheres.

Assim como nos países em desenvolvimento o número de mortes por câncer de colo do útero continua alto. Estudos históricos mostram que a taxa de mortalidade não varia desde 1985 (Camara *et al.*, 2003).

Na figura 2 pode-se observar o número médio de anos potenciais de vida perdidos devido ao câncer cervical. A faixa etária com maior taxa é a mesma que tem o maior potencial de ter o câncer. Essa taxa é ainda maior se o limite superior for de 80 anos (Inca, 2002).



Figura 2. Taxa do número médio de anos potenciais perdidos por câncer do colo do útero (Inca, 2002).

3. CÂNCER DO COLO DO ÚTERO

3.1. Fatores de Risco

Estudos epidemiológicos associaram a progressão de lesões de baixo risco ao câncer do colo do útero a vários fatores de risco. O principal fator associado a este câncer é a detecção da presença do Papillomavirus Humano (HPV) (de Villiers *et al.*, 2004; Kotecha *et al.*, 2003; Munoz *et al.*, 2003; zur Hausen, 2002; zur Hausen, 2001; Munoz, 2000; zur Hausen, 2000).

A correlação entre infecções genitais por HPV e o câncer do colo do útero foi primeiramente demonstrada pelo virologista alemão zur Hausen nos anos de 1980 (Burd, 2003). Em 1995/1996 a Organização Mundial da Saúde (OMS) reconheceu a infecção por HPV tipos 16 e 18 como “carcinogênica”, isto é, um fator importante na formação do câncer cervical (Burd, 2003; Munoz *et al.*, 2003; Franco *et al.*, 2001).

Outros fatores como: tabagismo; fatores hormonais (endógenos e exógenos); número de filhos; uso de contraceptivo oral; obesidade; coinfeção com outras doenças sexualmente transmissíveis como HSV-2 e *Chlamydia trachomatis*; número de parceiros sexuais; atividade sexual; idade da iniciação sexual; comportamento sexual da mulher e do parceiro; nutrição e polimorfismo na proteína p53 também estão associados em maior ou menor grau ao desenvolvimento do câncer (Castellsague *et al.*, 2006; Slomovitz *et al.*, 2006; Burd, 2003; de Araujo Souza e Villa, 2003; Franco *et al.*, 2001).

Um outro fator de risco importante e que está sendo analisado é a exposição ao sol das mulheres portadoras de HPV. Hrushesky (2006) analisou 900.000 testes de Papanicolau obtidos durante 16 anos e confrontou estes dados com a incidência dos raios UVB deste mesmo período. Os resultados ressaltaram um pico no verão para

os testes de Papanicolau que possuíam alterações epiteliais da cérvix. Neste mesmo período, o gráfico da incidência dos raios UVB apresentou o mesmo pico. Hrushesky concluiu que a incidência dos raios UVB pode ser um fator de risco importante para a progressão do câncer do colo do útero.

3.2. Epidemiologia

A transmissão do HPV ocorre primariamente do contato da pele nas relações sexuais. O HPV é resistente ao calor e a dessecação, podendo ocorrer transmissões não sexuais (Burd, 2003). Através de micro lesões o vírus entra no epitélio e infecta as células basais onde mantém entre 50 e 100 cópias do genoma por célula (McMurray *et al.*, 2001).

Uma simulação feita por Burchell *et al.* (2006) sugere que a transmissibilidade do HPV é maior que outros vírus sexualmente transmissíveis como o HIV e o HSV2.

Infecções clínicas e subclínicas por HPV, são hoje, as mais comuns das Infecções Sexualmente Transmissíveis (ISTs) tanto em homens quanto em mulheres (Baseman e Koutsky, 2005; Burd, 2003). Infecções assintomáticas podem ser detectadas entre 5 a 40% das mulheres em idade reprodutiva (Burd, 2003; Franco *et al.*, 2001).

Estas infecções são em sua maioria eliminadas pelo sistema imunológico, não deixando evidência da presença do HPV e não apresentando manifestações clínicas (Baseman e Koutsky, 2005).

A maioria das infecções por HPV é benigna e estão associadas a verrugas nas mãos e pés (Burd, 2003). Entretanto a associação entre a infecção por HPV e a

neoplasia cervical é alta. É mais alta que, por exemplo, a associação entre o cigarro e o câncer de pulmão e é somente comparada a associação entre hepatite B crônica e câncer de fígado (Franco *et al.*, 2001).

Em sua maioria o câncer do colo do útero acontece nas junções escamocolunar entre o epitélio colunar da endocérnix e o epitélio escamoso da ectocérnix. Neste local ocorrem mudanças metaplásicas freqüentes, que na puberdade e na primeira gravidez tem sua maior atividade, diminuindo com a chegada da menopausa. Em mulheres abaixo dos 40 anos o HPV está presente em 89% dos casos de adenocarcinomas e em mulheres acima dos 60 anos ou mais esta porcentagem é de 43% (Figura 3) (Burd, 2003).

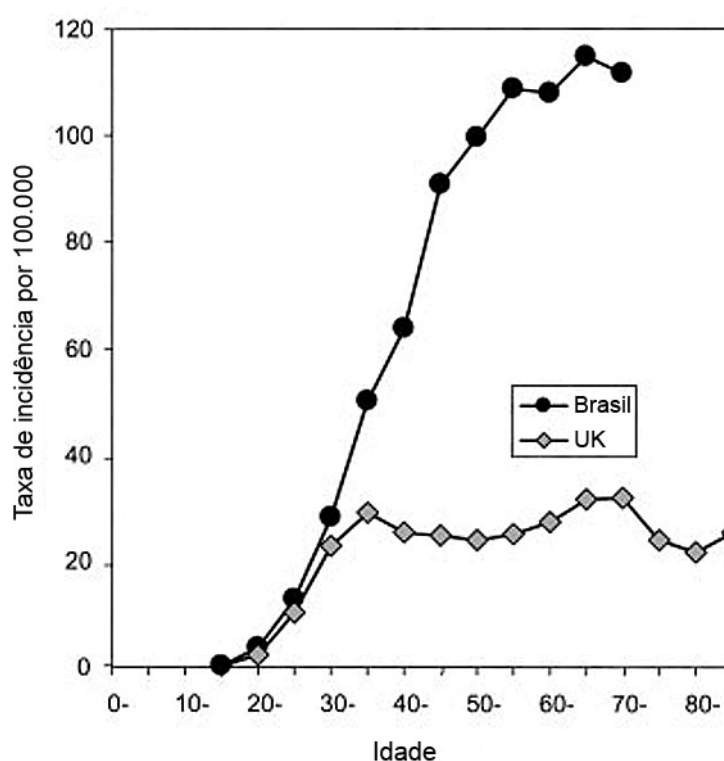


Figura 3. Taxa de incidência por grupo de idade para 100.000 mulheres para o Brasil e a Inglaterra. Adaptado de (Bosch e de Sanjose, 2003).

O câncer cervical é uma doença de progressão gradual. Acredita-se que primeiro existam neoplasias cervicais brandas (CIN1) para depois aparecer as mais severas neoplasias e lesões microinvasivas (CIN2 e CIN3) e ao final o câncer invasivo. É plausível aceitar que a infecção por HPV ocorra cedo, permanecendo latente na célula. Com a associação a outros fatores o desencadeamento do câncer cervical ocorre. Entretanto alguns pesquisadores estão sugerindo uma diferenciação entre o CIN1 e o CIN2 e CIN3. Estes últimos seriam os verdadeiros precursores do câncer do colo do útero. O risco de neoplasias brandas progredirem para neoplasias severas é de 1% ao ano, enquanto o risco das moderadas é de 16% em dois anos e 25% em cinco anos (Burd, 2003).

O risco subsequente de progressão a uma Neoplasia Cervical Intra-epitelial (CIN) é proporcional à positividade dos testes ao HPV, sugerindo desta forma que o desenvolvimento do carcinoma é resultado da persistência da infecção por este vírus (Franco *et al.*, 2001).

Com o uso da técnica de Reação em Cadeia de Polimerase (PCR) foi possível determinar que em 99,7% dos casos de câncer cervical o DNA do HPV estava presente. Este fato pode indicar que a infecção por HPV não é somente mais um dos fatores de risco, mas sim necessário para o desenvolvimento da neoplasia cervical (Castellsague *et al.*, 2006; Cruickshank, 2003; Franco *et al.*, 2001).

3.3. Diagnóstico

O principal método na detecção de alterações celulares no tecido do colo do útero é o teste de Papanicolau. Este teste foi desenvolvido pelo patologista George Papanicolaou em 1949, antes mesmo da causa do câncer cervical ser conhecida.

Desde então passou por modificações na sua classificação, para se adaptar aos novos conhecimentos sobre o câncer. (Burd, 2003).

Este teste tem uma especificidade alta, mas uma sensibilidade baixa. Segundo Oh *et al.* (2001) é impossível eliminar resultados falso-negativos e falso-positivos da citologia cervical.

Métodos de coleta e processamento para as amostras do Papanicolau foram recentemente desenvolvidos e aprovados para reduzir o número de falsos negativos, um dos problemas deste teste subjetivo (Burd, 2003).

Várias estratégias de protocolos de triagem para a presença do HPV e lesões cervicais podem ser utilizadas na tentativa de redução de custos (Legood *et al.*, 2006).

Diagnósticos baseados na detecção do DNA do HPV por PCR foram desenvolvidos para auxiliar o teste de Papanicolau (Burd, 2003). Atualmente o teste comercialmente mais usado e aceito é o *Híbrido Capture 2* (Brink *et al.*, 2006)

Estudos recentes demonstram que o uso desta metodologia é mais sensível e mais eficiente na detecção da presença do HPV e alterações associadas ao câncer do colo do útero do que o teste de Papanicolau (Castellsague *et al.*, 2006; Burd, 2003; Lorenzato *et al.*, 2001; Lorenzato *et al.*, 2000; Ratnam *et al.*, 2000).

Este método pode ser utilizado de três formas: A primeira na identificação de mulheres com risco de desenvolver o câncer cervical e em programas de triagem. A segunda em experimentos com vacinas e estudos epidemiológicos obtendo-se o máximo de informações sobre o HPV em determinada população (Molijn *et al.*, 2005). Por último a detecção do DNA do HPV pode ser utilizada após o tratamento para ajudar a identificar a presença ou ausência de lesões residuais (Burd, 2003).

3.4. Prevenção e tratamento

A prevenção é uma das melhores formas para o controle das infecções sexualmente transmissíveis (IST) como é o caso do HPV. Desta forma as mesmas atitudes tomadas em relação a outras doenças podem ajudar no combate à transmissão e infecção pelo HPV. A fidelidade a um único parceiro sexual diminui a probabilidade da infecção já que a multiplicidade de parceiros é um dos fatores de risco para esta doença (Tristram, 2006; Baseman e Koutsky, 2005).

A disponibilização de informações sobre o contágio, transmissão e as possíveis consequências levam à possibilidade de mudar o comportamento de risco (Tristram, 2006).

O teste preventivo de Papanicolau continua sendo o principal método de prevenção do câncer do colo do útero. Quando diagnosticadas displasias brandas ou moderadas em estágios menos avançados, apresentam alta taxa de regressão nos primeiros dois anos de diagnóstico (Brink *et al.*, 2006; Holowaty *et al.*, 1999).

A vacina profilática, que está em fase III, é outra forma de prevenção da infecção por HPV. Para os portadores do vírus uma outra abordagem é a vacina terapêutica. Estas vacinas estão baseadas nos genes E6, E7, L1 e L2 (Stern, 2005; Frazer, 2004; Frazer, 2002; Steller, 2002; McMurray *et al.*, 2001).

Muitas perguntas sobre estas vacinas ainda estão sem respostas. Quais os tipos de HPVs que devem ser incluídos? Qual a duração da vacina profilática? Será preciso diferentes coquetéis para diferentes populações? Quem deve ser imunizado, mulheres e/ou garotas? Os do sexo masculino também devem tomar a vacina? Muitas destas perguntas certamente só serão respondidas com o início da vacinação (Stanley, 2006).

4. PAPILLOMAVIRUS HUMANO (HPV)

Até o final dos anos 1970 os Papillomavirus atraíam pouco interesse dos pesquisadores por estarem então associados a verrugas casuais e lesões benignas da pele. Com o desenvolvimento de novas técnicas moleculares nos anos 1980 foi possível detectar muitos tipos de HPVs em lesões benignas e malignas nas mucosas (Bernard, 2005; zur Hausen, 2002).

Mais de 200 tipos de HPV são conhecidos, sendo 85 a 100 destes genomas totalmente seqüenciados e bem caracterizados. Outros 120 isolados foram parcialmente caracterizados e representam potenciais novos genomas (Bernard, 2005; Burd, 2003).

4.1. Estrutura

Os HPVs são vírus relativamente pequenos, não envelopados, assim não possuindo camadas lipídicas que o envolvem. Seu capsídio é composto por 72 capsômeros. Cada capsômero é formado por um pentâmero e cada capsídeo viral contém várias cópias, por volta de 12, da proteína estrutural. Possui uma forma icosaédrica ou poliédrica com 20 faces, com um tamanho de 55 nm de diâmetro (Figura 4) (Bernard, 2005; Greenblatt, 2005; Burd, 2003).

O genoma do HPV é uma molécula circular simples de dupla fita com aproximadamente 7.900 pares de bases (pb) O genoma é dividido em três regiões, e possuem 8 códigos abertos de leitura (ORF), alguns destes sobrepostos, que codificam 10 proteínas.

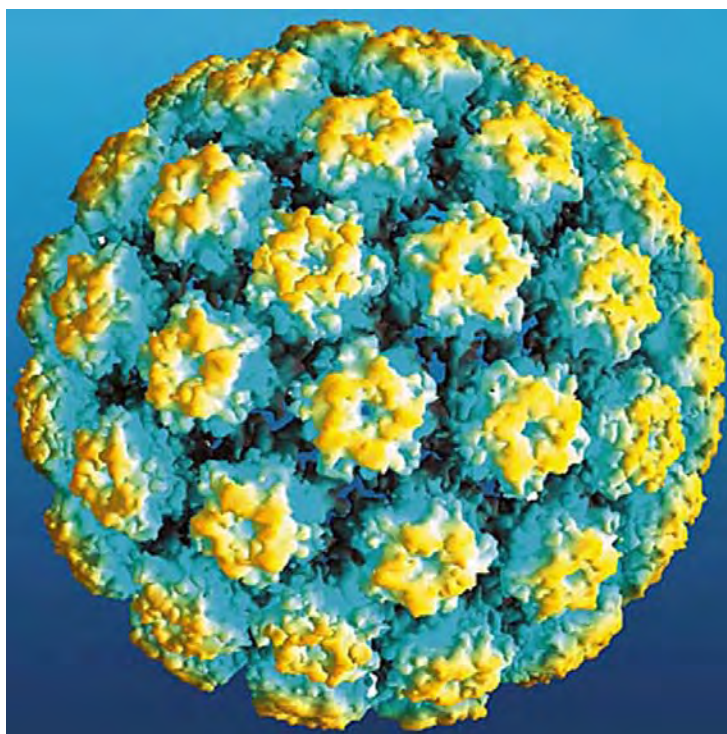
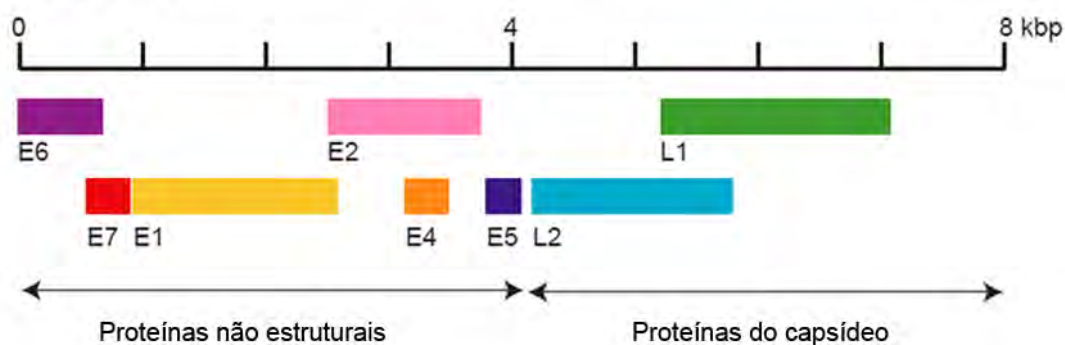


Figura 4. A ilustração mostra uma vista da superfície molecular do modelo atômico do papillomavirus proposta por (Modis *et al.*, 2002).

A primeira região é não codificante, 400 a 1000 pb, e é chamada de Região Longa de Controle (LCR), e apresenta a maior variação do genoma viral. A segunda região chamada de precoce compreende as proteínas: E1, E2, E3, E4, E5, E6, E7, E8 que estão envolvidas na replicação viral. A terceira região, tardia, é formada pelas proteínas L1 e L2, proteínas estruturais do capsídeo viral ([Figura 5](#)) (Greenblatt, 2005; Molijn *et al.*, 2005; Motoyama *et al.*, 2004; Munger *et al.*, 2004; Burd, 2003; Frazer, 2002).

Os genes E1, E2 e E4 estão envolvidos na replicação, transcrição e maturação viral. E5, E6, e E7 são oncoproteínas envolvidas na transformação celular. Os genes L1 e L2 codificam as proteínas estruturais do capsídeo viral (de Villiers *et al.*, 2004; Hilleman, 2000).

A - HPV 16



B - HPV 11

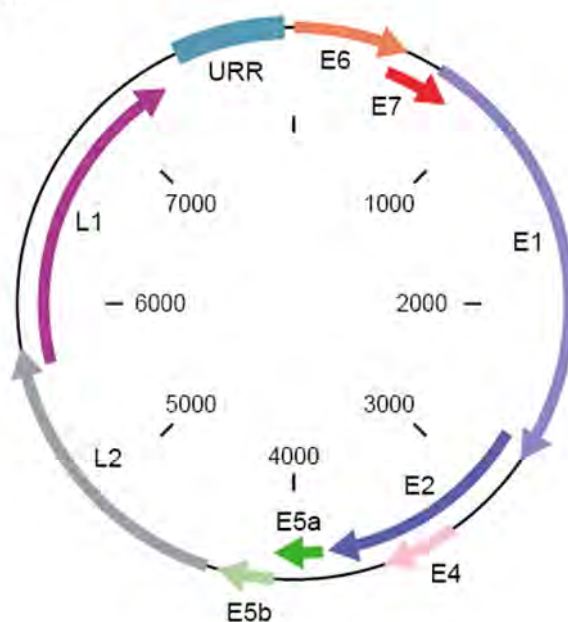


Figura 5. Ilustrações de dois tipos de Papillomavirus humano. A - Organização simplificada do genoma do HPV-16 (Linearizado). B - Organização circular do genoma do HPV 11. Os E genes codificam as proteínas não estruturais. Os L genes codificam as duas proteínas do capsídeo viral.

4.2. Patogenicidade

As proteínas E6 e E7 possuem um pequeno tamanho molecular. São capazes de interagir com toda a variedade de proteínas reguladoras da célula hospedeira, e desta forma levar a imortalização celular e ao desenvolvimento do processo neoplásico (Finzer *et al.*, 2002).

Além disso, quando associadas, desregulam o mecanismo de duplicação dos cromossomos e a segregação durante a mitose e induz a instabilidade cromossomal (Wentzensen *et al.*, 2004; Duensing *et al.*, 2000).

A proteína E6 possui 151 aminoácidos contendo dois domínios de dedos de zinco. Sozinha não é capaz de immortalizar a célula humana, mas provê várias funções que alteram o meio celular. Estas alterações incluem o bloqueio da apoptose através da degradação do p53, alteração dos genes da transcrição celular através da interação com a p300, e o aumento da atividade telomerásica. Estas alterações são necessárias para facilitar o ciclo de vida do vírus (McMurray *et al.*, 2001).

Após a ligação com a proteína E6-associada (E6AP), uma proteína ligase da via de proteólise da ubiquitina, o complexo se dirige a proteína p53, proteína supressora de tumor e de regulação do ciclo celular, inativando-a e levando a célula a instabilidade genética e, desta forma ao desenvolvimento de um quadro maligno evolutivo, causado pela perda da função da p53 na regulação da célula ao estresse e aos danos genômicos (McMurray *et al.*, 2001).

O gene E7 codifica uma proteína de 98 aminoácidos, que possui na sua região final um motivo de dedos de zinco, essencial para a dobradura da proteína (McMurray *et al.*, 2001).

A proteína E7 se liga preferencialmente à proteína retinoblastona (Rb), rompendo o complexo Rb-E2F, um dos fatores de transcrição e causando um efeito cascata destes fatores, o que por sua vez leva ao aumento da expressão dos genes envolvidos no avanço do ciclo celular e na síntese de DNA (Figura 6) (Burd, 2003; Finzer *et al.*, 2002; Munger e Halpern, 1997).

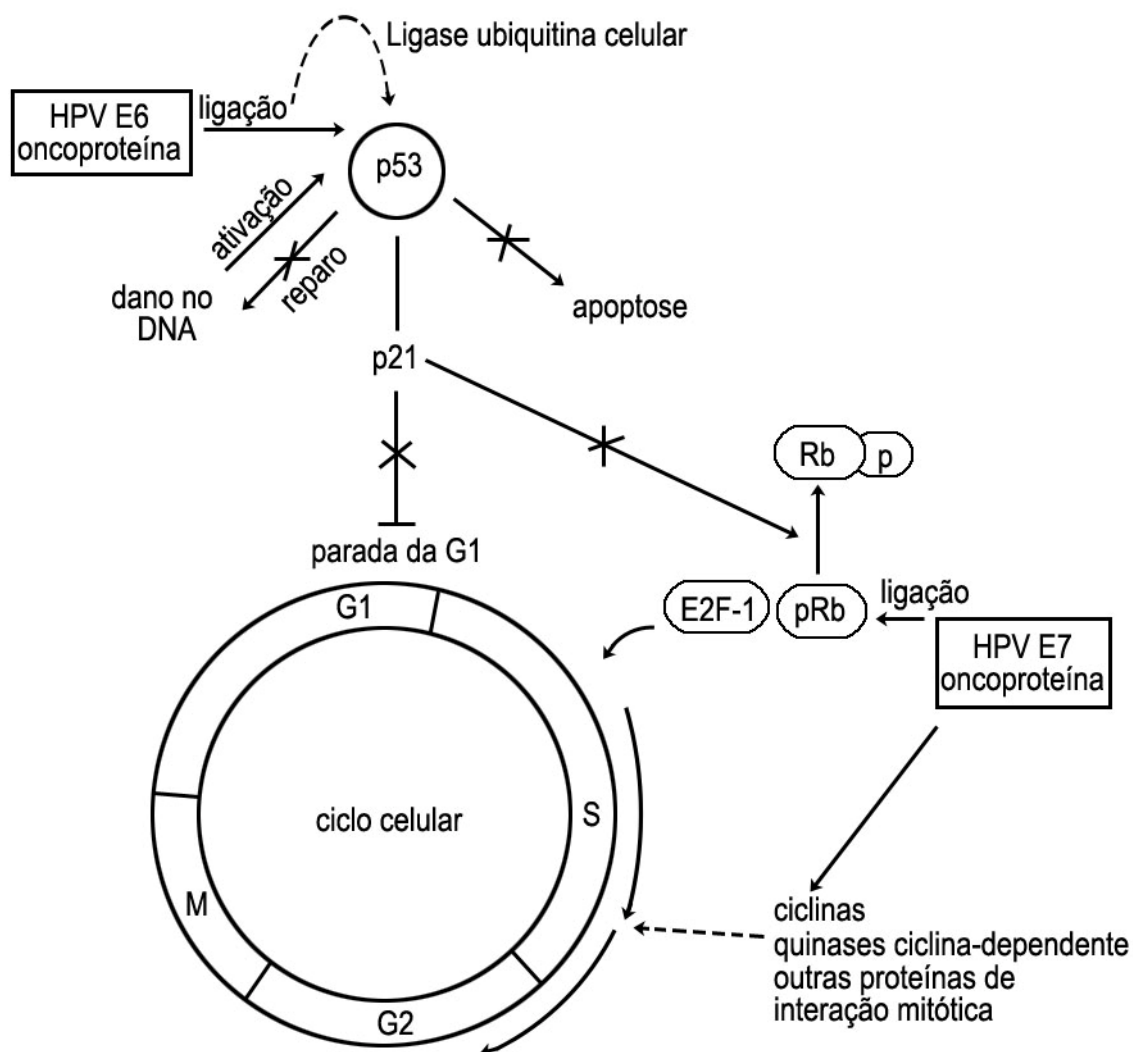


Figura 6. Patogenicidade do HPV. Os genes E6 e E7 codificam proteínas multifuncionais que se ligam primariamente às proteínas p53 e pRb, alterando o ciclo celular. Modificado de (Burd, 2003).

A afinidade entre a E7 e a pRb varia com o tipo de HPV. O HPV-6, por exemplo, tem uma baixa afinidade no acoplamento entre as duas proteínas. Isso se deve provavelmente às mudanças de aminoácidos perto do sítio de ligação LxCxE. Já a proteína E7 do HPV-16 tem uma alta afinidade na sua ligação com a proteína pRb (McMurray *et al.*, 2001).

A proteína E5 desestabiliza a membrana celular e modifica a resposta celular para a morte da célula ou apoptose. A proteína E4 se liga às proteínas do citoesqueleto e impede uma diferenciação celular apropriada (Greenblatt, 2005; McMurray *et al.*, 2001)

O produto gênico do gene E2 é uma proteína de ligação ao DNA que bloqueia a transcrição dos genes E6 e E7 permitindo que o produto do gene E1 se ligue a origem de replicação viral localizada na LCR. Esta ligação inicia a replicação do genoma viral na fase S (Burd, 2003).

Análises de microarranjos de células infectadas pelo HPV-31 mostraram que 178 genes celulares tiveram sua expressão aumentada e 150 outros genes tiveram sua expressão diminuída. Os genes que sofreram baixas na sua expressão estão envolvidos na regulação do crescimento celular (Burd, 2003)

4.3. Classificação

Os Papillomavirus são altamente diversificados e ocorrem na maioria dos mamíferos e aves (de Villiers *et al.*, 2004). Pertencem a um grupo de vírus que são agentes etiológicos de lesões epitélio cutâneo e mucosas que leva à formação de tumores (Kotecha *et al.*, 2003).

Recentemente o Conselho Internacional em Taxonomia dos Vírus (ICTV) reconheceu e oficializou a criação da família *Papillomaviridae*, onde agora se encontram os Papillomavirus Humanos (Bernard, 2005; de Villiers *et al.*, 2004).

Como resultado de esforços de um grande número de seqüenciamentos dos tipos de HPV, esta família possui hoje o DNA mais conhecido entre as famílias dos

vírus e com isso é uma das mais importantes nos estudos de evolução viral (Chan *et al.*, 1995).

4.3.1. Filogenia

A validação entre tipos, subtipos e variantes é determinada pelas diferenças no código genético, os nucleotídeos, da proteína L1. A ORF L1 é o gene mais conservado do genoma dos Papillomavirus e por isso vem sendo usada nos últimos 15 anos para sua identificação e classificação (Figura 7) (de Villiers *et al.*, 2004).

Para que um novo tipo de HPV seja registrado é necessário que seu genoma tenha no mínimo 10% de diferença entre todos os HPVs conhecidos. Foi criado o termo subtipos para diferenças na L1 abaixo de 10% e acima de 2% (HPV-6a, HPV-6b). Alguns HPVs anteriormente às técnicas atuais de biologia molecular foram classificados como tipos isolados e hoje permanecem com seus números apesar de serem subtipos. Os tipos 46, 55, 64 são hoje subtipos dos 20, 44, 34 respectivamente. Diferenças abaixo de 2% entre o genoma novo e o protótipo, genoma de referência, são classificados como variantes. Estes variantes têm sua diferença máxima no genoma quando se compara grupos étnicos com uma longa separação temporal (Bernard, 2005; de Villiers *et al.*, 2004).

Entre os HPVs, o tipo 16 foi o mais seqüenciado e melhor teve caracterizado seus variantes. Foram encontrados 5 variantes principais: Europeu (E), Asiático (As), Asiático-Americano (AA), Africano 1 (Af-1) e Africano 2 (Af-2). Estas variações nucleotídicas em um gene estão normalmente associadas a mudanças em outro gene sugerindo que os variantes são estáveis e não estão sujeitos a recombinações freqüentes (Swan *et al.*, 2005).

Os Papillomavirus são classificados em 16 gêneros diferentes. Cinco destes gêneros têm como hospedeiro o homem: alfa-papillomavirus, beta-papillomavirus, gama-papillomavirus, Mu-papillomavirus e Nu-papillomavirus (de Villiers *et al.*, 2004). O gênero clinicamente mais importante é o alfa-papillomavirus. Nele estão presentes todos os tipos de HPV associados a lesões das mucosas e genitais. O termo alfa-papillomavirus torna-se mais neutro que HPVs da mucosa e HPVs genitais anteriormente usados. Exemplo disto são os tipos 6 e 11 classificados como papillomas genitais por serem comumente encontrados nas verrugas genitais, mas que foram encontrados também na laringe (Bernard, 2005).

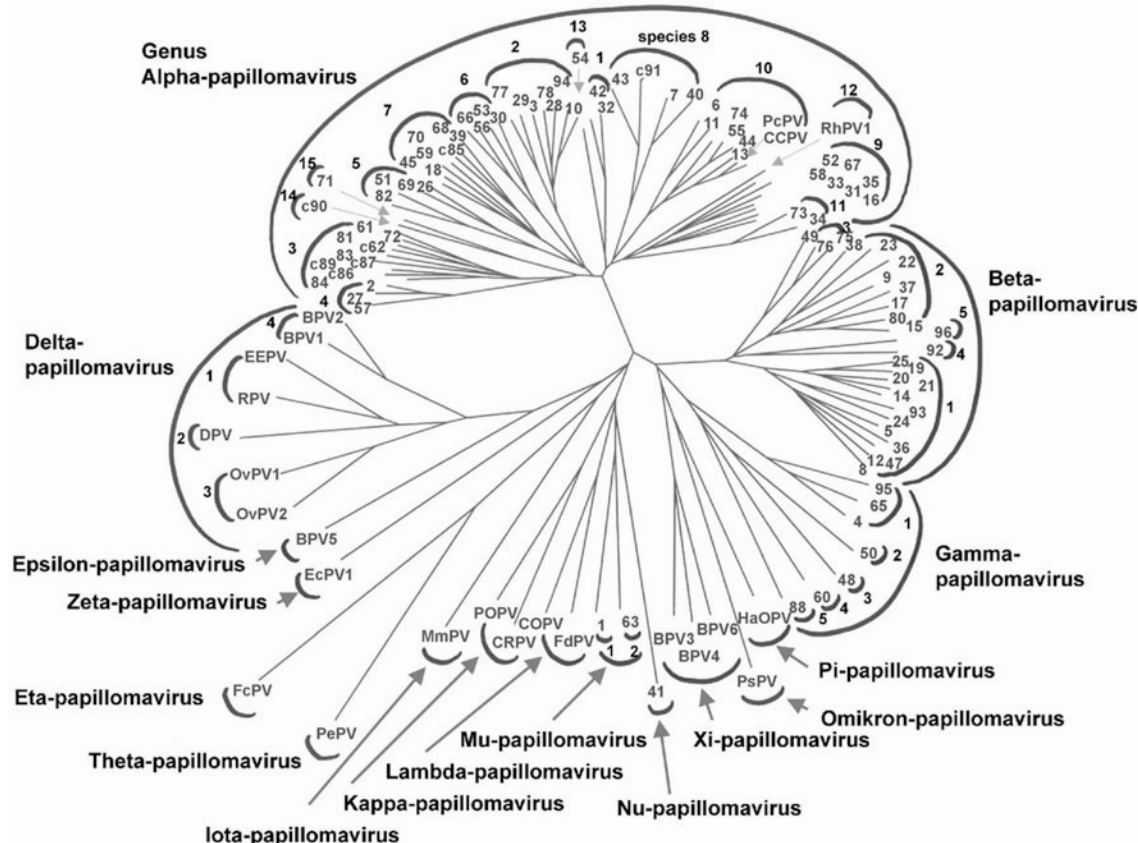


Figura 7. Árvore filogenética montada a partir da sequência do gene L1 de 118 tipos de Papillomavirus humano (de Villiers *et al.*, 2004).

Acredita-se que a evolução molecular dos tipos de HPV ocorra por acumulação gradual e mutações simples. Para os HPVs 16 e 33, tipos bem próximos, este modelo de evolução molecular parece indicar uma variação genômica contínua, que se estende da época do antepassado comum até o mais recente. Pôde-se assim identificar dentre variantes aqueles que são os mais próximos de seu antepassado comum (Chan *et al.*, 1992).

4.3.2. Alto e baixo risco associado ao câncer cervical

Os Papillomavirus humanos são classificados epidemiologicamente em alto risco e baixo risco quanto a sua associação com o câncer do colo do útero e lesões precursoras para câncer. Os HPVs de baixo risco estão associados a inflamações cutâneas e verrugas e são benignos, mesmo que ocasionalmente encontrados em carcinomas associados aos de alto risco. Já os HPVs de alto risco estão associados diretamente a lesões precursoras e com carcinoma invasivo (Burd, 2003; Munoz *et al.*, 2003). A tabela 4 sintetiza os HPVs classificados em alto e baixo risco para os tipos que foram utilizados neste trabalho (Bernard, 2005; de Villiers *et al.*, 2004; Bosch e de Sanjose, 2003; Burd, 2003; Munoz *et al.*, 2003; Lorenzato *et al.*, 2001; Matsukura e Sugase, 2001; McMurray *et al.*, 2001; Lorenzato *et al.*, 2000; Munoz, 2000; de Villiers, 1989).

Divergências entre a classificação filogenética e a epidemiológica ocorreram em dois tipos de HPV. O HPV-70 que filogeneticamente foi classificado como de alto risco e o tipo 73 classificado como baixo risco tem a classificação invertida, baixo e alto risco respectivamente na classificação epidemiológica (Munoz *et al.*, 2003).

Tabela 4. Classificação epidemiológica dos tipos de Papillomavirus Humano associada ao alto e baixo risco para o câncer do colo do útero.

Tipo	Risco	Tipo	Risco	Tipo	Risco	Tipo	Risco
HPV-1*	Baixo	HPV-22	Baixo	HPV-47	Baixo	HPV-72*	Baixo
HPV-02*	Baixo	HPV-23	Baixo	HPV-48	Baixo	HPV-73	Baixo/Alto
HPV-02a	Baixo	HPV-24	Baixo	HPV-49	Baixo	HPV-74	Baixo
HPV-3*	Baixo	HPV-25	Baixo	HPV-50	Baixo	HPV-75	Baixo
HPV-04	Baixo	HPV-26	Alto	HPV-51	Alto	HPV-76	Baixo
HPV-05	Baixo	HPV-27	Baixo	HPV-52	Alto	HPV-77	Baixo
HPV-06	Baixo	HPV-28	Baixo	HPV-53*	Alto	HPV-78*	Baixo
HPV-06a	Baixo	HPV-29	Baixo	HPV-54	Baixo	HPV-80	Baixo
HPV-06b	Baixo	HPV-30	Alto	HPV-55	Baixo	HPV-81	Baixo
HPV-07	Baixo	HPV-31	Alto	HPV-56	Alto	HPV-82	Alto
HPV-08	Baixo	HPV-32	Baixo	HPV-57	Baixo	HPV-83	Baixo
HPV-09	Baixo	HPV-33	Alto	HPV-58	Alto	HPV-84	Baixo
HPV-10	Baixo	HPV-34	Alto	HPV-59	Alto	HPV-85	Alto
HPV-11	Baixo	HPV-35	Alto	HPV-60	Baixo	HPV-86	Baixo
HPV-12	Baixo	HPV-36	Baixo	HPV-61	Baixo	HPV-87	Baixo
HPV-13	Baixo	HPV-37	Baixo	HPV-62	Baixo	HPV-88*	Baixo
HPV-14*	Baixo	HPV-38	Baixo	HPV-63	Baixo	HPV-89	Baixo
HPV-15	Baixo	HPV-39	Alto	HPV-65	Baixo	HPV-90	Baixo
HPV-16	Alto	HPV-40	Baixo	HPV-66	Alto	HPV-91	Baixo
HPV-17	Baixo	HPV-41*	Baixo	HPV-67	Alto	HPV-92	Baixo
HPV-18	Alto	HPV-42*	Baixo	HPV-68*	Alto	HPV-93*	Baixo
HPV-19	Baixo	HPV-43	Baixo	HPV-69*	Alto	HPV-94*	Baixo
HPV-20	Baixo	HPV-44	Baixo	HPV-70*	Alto/Baixo	HPV-95*	Baixo
HPV-21	Baixo	HPV-45	Alto	HPV-71*	Baixo	HPV-96*	Alto

* Tipos de HPVs não utilizados neste trabalho por não possuir as exigências necessárias.

4.4. Os tipos mais comuns associados à oncogenicidade

O HPV tipo 16 é o mais comum em mulheres citologicamente normais, e é também o mais comum dos tipos associados a casos de câncer do colo do útero (Baseman e Koutsky, 2005; Munoz *et al.*, 2003; Franco *et al.*, 1999). Este foi encontrado em 50 a 60% dos casos de câncer cervical, seguido do HPV 18 com 10 a 15% e do tipo 45 com 5% e o HPV 33 com 4% (Bosch e de Sanjose, 2003; Munoz *et al.*, 2003; Hildesheim *et al.*, 2001). Os tipos 31 e 35 são mais comuns na América Latina que nas outras áreas estudadas (Figura 8) (Munoz *et al.*, 2003).

Entre os tipos mais comuns, HPV 16 e 18, foram encontradas diferenças na oncogenicidade entre seus variantes (Arias-Pulido *et al.*, 2005; Yamada *et al.*, 1997).

Em estudo desenvolvido no Brasil, Villa *et al.* (2000) encontraram sete variantes moleculares para quatro dos variantes geográficos. Para o tipo 16, protótipo Europeu, foi encontrado em 54% dos casos estudados, seguido pelo variante Asiático-Americano com 22%. Já para o tipo 18 foram encontrados 4 variantes moleculares classificados como protótipo Europeu e somente um caso para o Africano.

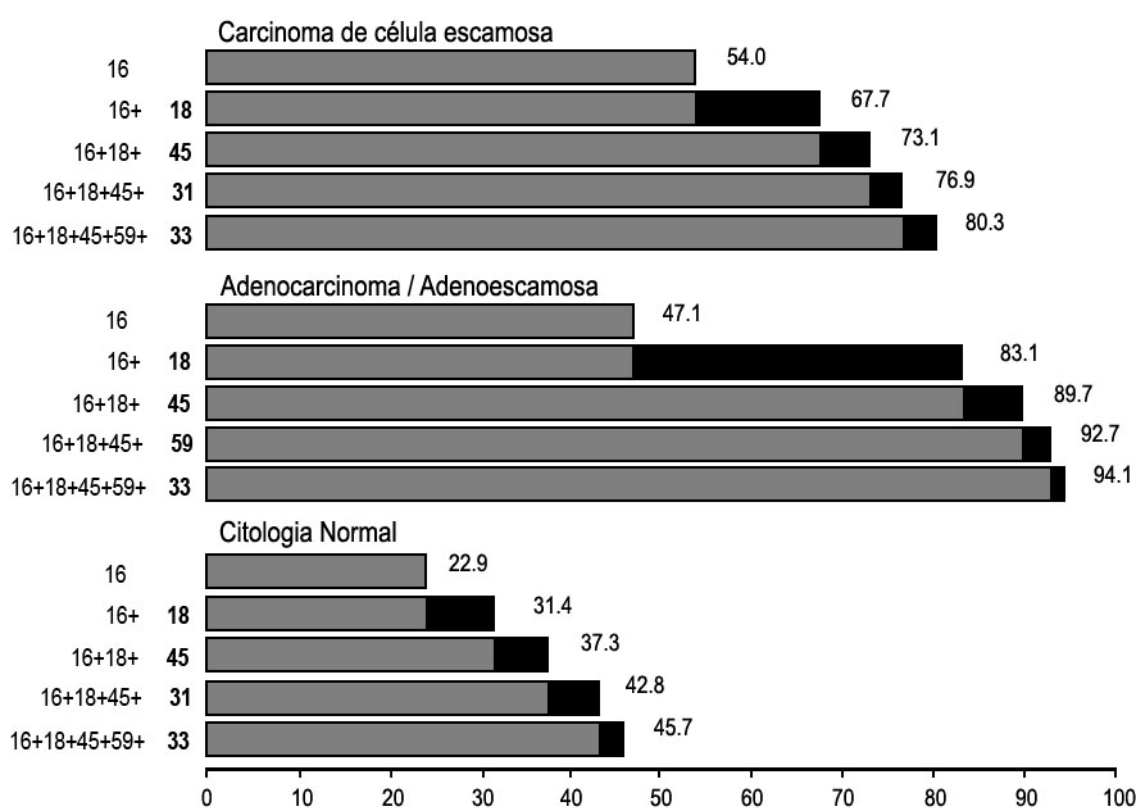


Figura 8. Prevalência cumulativa dos tipos mais comuns de HPVs associados a casos de câncer cervical. Adaptado de (Bosch e de Sanjose, 2003).

Hildesheim (2001) encontrou associações diferentes com a neoplasia cervical para os variantes do HPV tipo 16. O autor chegou a uma diferença de 11 vezes no poder carcinogênico do variante não europeu quando comparado com o variante europeu.

5. ABORDAGEM *IN SILICO*

Com o surgimento dos seqüenciadores automáticos, e de novas e eficientes técnicas na análise de seqüências de genomas, os atuais bancos de dados biológicos crescem exponencialmente (Prosdocimi *et al.*, 2003; Souto *et al.*, 2003).

Esta elevada quantidade de informação exige recursos computacionais cada vez mais eficientes para o armazenamento e análise destes dados (Prosdocimi *et al.*, 2003; Lesk, 2002).

Por outro lado, esta disponibilidade de grandes quantidades de genomas seqüenciados, de muitas espécies diferentes, oferece novas oportunidades de estudos (Yu *et al.*, 2004).

5.1. Bioinformática – Ciência multidisciplinar na busca de informações

Com a utilização de técnicas e ferramentas computacionais para a resolução de problemas em biologia, e em especial a Biologia Molecular, surgiu a Bioinformática ou Biologia Computacional (Souto *et al.*, 2003).

Esta nova ciência envolve linhas diversas de conhecimento, como a matemática, a estatística, a ciência da computação e a biologia molecular. Biólogos moleculares passaram a utilizar métodos estatísticos para analisar grandes quantidades de dados biológicos (Eddy, 2005; Barnes e Gray, 2003; Prosdocimi *et al.*, 2003; Lesk, 2002).

A bioinformática pode ser aplicada na resolução de vários problemas biológicos como: reconhecimento de genes, identificação e análise da expressão gênica, determinação de estruturas protéicas, e identificação de possíveis diferenças

nas seqüências que podem favorecer o desenvolvimento de doenças (Barnes e Gray, 2003; Souto *et al.*, 2003).

As informações sobre a similaridade entre genomas de organismos é um dos recursos principais na busca por regiões funcionais e também em prever a ação destas regiões (Chang, 2005; Hardison, 2003). A análise de seqüências assume então grande importância para correlação entre organismos no nível molecular (Hardison, 2003).

Pesquisas sobre análise de seqüências, genômica comparativa e modelagem de proteínas, que utilizam a bioinformática, - com técnicas de aprendizado de máquina -, têm sido amplamente utilizadas para estudos do genoma e proteoma. Pode-se destacar entre estes estudos: identificação de regiões gênicas sobrepostas no genoma viral (McCauley e Hein, 2006); predição de ilhas genômicas – transferência horizontal de genes – em procariontes (Waack *et al.*, 2006); efeitos biológicos de mutações na p53 comparando-se seqüências (Mathe *et al.*, 2006); classificação dos vírus da família HPV entre alto e baixo risco através da sua seqüência de aminoácidos (Joung *et al.*, 2005; Joung *et al.*, 2004); triagem virtual de compostos com possível atividade biológica para alvos terapêuticos (Jorissen e Gilson, 2005); análise comparativa dos genes cortactin e HSI relacionados a carcinomas em humanos (van Rossum *et al.*, 2005); busca por genes dentro do genoma através de ferramentas matemáticas (Majoros *et al.*, 2005; Korf, 2004; Pedersen e Hein, 2003; Henderson *et al.*, 1997); construção de *primers* por algoritmos genéticos (Wu *et al.*, 2004); detecção de homologia entre classes de proteínas (Leslie *et al.*, 2004); identificação de elementos conservados no genoma (Lenhard *et al.*, 2003); análise dos genes de bactérias *E. coli* patogênicas e não-patogênicas para encontrar possíveis diferenças nucleotídicas (Perna *et al.*, 2001);

identificação de genes relacionados a doenças humanas (Jimenez-Sanchez *et al.*, 2001); classificação e validação de dados de microarranjos de amostras de tecidos cancerígenos (Furey *et al.*, 2000) e alinhamento e modelagem de famílias de proteínas com base na sua estrutura primária (Baldi *et al.*, 1994).

5.2. Modelos Ocultos de Markov – HMM

Análises comparativas de seqüências de nucleotídeos ou aminoácidos são uma das ferramentas principais na elucidação de mecanismos moleculares de replicação e evolução dos organismos e das estruturas e funções das proteínas (Eddy, 2005; Baldi *et al.*, 1994).

No alinhamento de seqüências busca-se associar a seqüência alvo com as seqüências constantes do banco de dados. Um dos métodos mais usados para estes alinhamentos são os Modelos Ocultos de Markov (HMM) (Joung *et al.*, 2005; Eddy, 2004).

Eddy (2004) os chama de “os Legos” da análise computacional de seqüência, pois podem ser utilizados em diversas abordagens: buscadores de genes, alinhamento múltiplo de seqüências, identificação de sítios de regulação.

Os HMM são uma classe de modelos estatísticos que podem ser aplicados na análise de seqüências de aminoácidos, pois conseguem tratar de forma probabilística a variação estrutural dos elementos de uma mesma classe biológica (Joung *et al.*, 2005; Souto *et al.*, 2003; Mount, 2001).

Estes modelos são amplamente utilizados no alinhamento múltiplo de seqüências como “perfis probabilísticos”, onde um subconjunto é utilizado para o treinamento do modelo. Para se chegar ao melhor alinhamento possível entre as

seqüências a frequência de cada aminoácido é contada e assim obtém-se a probabilidade associada a cada modelo (Mount, 2001).

5.3. Aprendizado de Máquina – ML

Técnicas de Aprendizado de Máquina (ML) são cada vez mais usadas como alternativa a modelos computacionais convencionais por trabalhar de forma eficiente com grandes quantidades de dados. Estas técnicas possuem a característica de “aprender” e melhorar automaticamente à medida que “aprendem” com experiências passadas e assim produzir hipóteses úteis e têm se mostrado uma alternativa promissora para a classificação de dados do genoma (Baldi e Brunak, 2001).

Estas técnicas podem ser divididas em aprendizado supervisionado e não supervisionado. Entre as técnicas supervisionadas estão as redes neurais artificiais, árvores de decisão, algoritmos genéticos e máquinas de vetores de suporte (Souto *et al.*, 2003).

Ferramentas de ML baseadas em Máquinas de Vetores de Suporte (SVM) têm se mostrado de grande utilidade na construção de classificadores com base em padrões identificados em cadeias de caracteres que representam a constituição química do DNA ou proteínas (Souto *et al.*, 2003; Baldi e Brunak, 2001).

O princípio desta técnica está em encontrar um hiperplano ótimo que separa membros e não-membros de uma classe em um espaço abstrato. As classes presentes no conjunto de treinamento se tornam linearmente separáveis. O hiperplano ótimo é definido como aquele para o qual a margem de separação entre as classes é maximizada (Figura 9) (Jorissen e Gilson, 2005; Souto *et al.*, 2003; Baldi e Brunak, 2001).

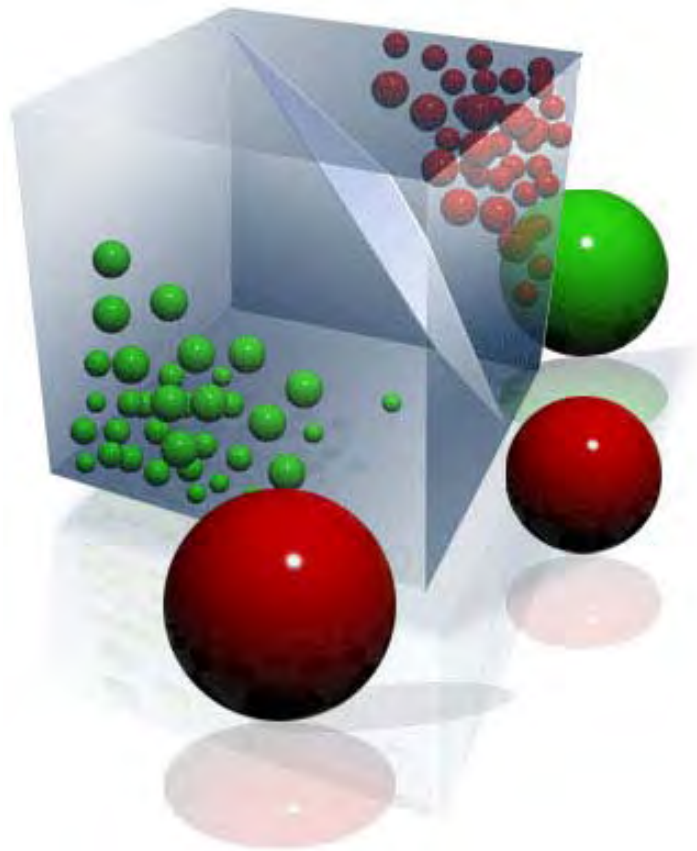


Figura 9. Representação do hiperplano ótimo, margem maximizada, do SVM. Dentro do cubo representantes das duas classes (vermelho e verde) separados nas margens opostas.

Muitos dos problemas computacionais precisam procurar por um número enorme de soluções. Exemplo disto é a busca por propriedades específicas de uma classe de proteínas, onde é necessário que o algoritmo procure num número vasto de seqüências de aminoácidos por uma solução possível. Estes problemas precisam de um programa de computador adaptável e que continue sua execução em um ambiente com mudanças (Melanie, 1999).

Para resolver estes problemas onde é necessário adaptação foi buscada inspiração na biologia e sua constante evolução nos processos biológicos, e foram desenvolvido os Algoritmos Genéticos (GAs) (Whitley, 1994).

Os GAs fazem parte da área do aprendizado de máquina conhecida como computação evolutiva, que estuda algoritmos inspirados na teoria da evolução e na genética (Souto *et al.*, 2003). Os GAs são algoritmos de busca e otimização utilizados para gerar variações entre soluções e combinar as características daquelas que obtiveram melhor desempenho, gerando novas soluções com desempenho melhor a cada geração (Baldi e Brunak, 2001).

Os GAs atuam sobre uma população de indivíduos, baseados no fato de que indivíduos com boas características genéticas têm maior chance de sobrevivência e de produzirem indivíduos cada vez mais aptos. Cada indivíduo recebe uma nota, refletindo a qualidade da sua solução para o problema em questão (Souto *et al.*, 2003; Baldi e Brunak, 2001).

Devido a estas características os GAs estão sendo cada vez mais utilizados na biológica computacional para modelagem de sistemas naturais (Melanie, 1999).

Neste projeto utilizamos estas três ferramentas – Modelos Ocultos de Markov, Máquinas de Vetores de Suporte e Algoritmos Genéticos – para classificar o HPV entre alto e baixo risco com base na cadeia de aminoácidos das proteínas.

REFERÊNCIAS BIBLIOGRÁFICAS

Arias-Pulido, H., Peyton, C. L., *et al.* Human papillomavirus type 18 variant lineages in United States populations characterized by sequence analysis of LCR-E6, E2, and L1 regions. Virology, v.338, n.1, Jul 20, p.22-34. 2005.

Baldi, P. e Brunak, S. Bioinformatics: The Machine Learning Approach. Cambridge: MIT Press. 2001. 477 p.

Baldi, P., Chauvin, Y., *et al.* Hidden Markov models of biological primary sequence information. Proc Natl Acad Sci U S A, v.91, n.3, Feb 1, p.1059-63. 1994.

Barnes, M. R. e Gray, I. C. Bioinformatics for Geneticists. England: John Wiley & Sons Ltd,. 2003. 423 p.

Baseman, J. G. e Koutsky, L. A. The epidemiology of human papillomavirus infections. J Clin Virol, v.32 Suppl 1, Mar, p.S16-24. 2005.

Bernard, H. U. The clinical importance of the nomenclature, evolution and taxonomy of human papillomaviruses. J Clin Virol, v.32 Suppl 1, Mar, p.S1-6. 2005.

Bosch, F. X. e De Sanjose, S. Chapter 1: Human papillomavirus and cervical cancer-burden and assessment of causality. J Natl Cancer Inst Monogr, n.31, p.3-13. 2003.

Brink, A. A., Snijders, P. J., *et al.* HPV testing in cervical screening. Best Pract Res Clin Obstet Gynaecol, v.20, n.2, Apr, p.253-66. 2006.

Burchell, A. N., Richardson, H., *et al*. Modeling the sexual transmissibility of human papillomavirus infection using stochastic computer simulation and empirical data from a cohort study of young women in Montreal, Canada. Am J Epidemiol, v.163, n.6, Mar 15, p.534-43. 2006.

Burd, E. M. Human papillomavirus and cervical cancer. Clin Microbiol Rev, v.16, n.1, Jan, p.1-17. 2003.

Camara, G. N., Cerqueira, D. M., *et al*. Prevalence of human papillomavirus types in women with pre-neoplastic and neoplastic cervical lesions in the Federal District of Brazil. Mem Inst Oswaldo Cruz, v.98, n.7, Oct, p.879-83. 2003.

Castellsague, X., Diaz, M., *et al*. Worldwide human papillomavirus etiology of cervical adenocarcinoma and its cofactors: implications for screening and prevention. J Natl Cancer Inst, v.98, n.5, Mar 1, p.303-15. 2006.

Chan, S. Y., Bernard, H. U., *et al*. Phylogenetic analysis of 48 papillomavirus types and 28 subtypes and variants: a showcase for the molecular evolution of DNA viruses. J Virol, v.66, n.10, Oct, p.5714-25. 1992.

Chan, S. Y., Delius, H., *et al*. Analysis of genomic sequences of 95 papillomavirus types: uniting typing, phylogeny, and taxonomy. J Virol, v.69, n.5, May, p.3074-83. 1995.

Chang, P. L. Clinical bioinformatics. Chang Gung Med J, v.28, n.4, Apr, p.201-11. 2005.

Cruickshank, M. E. The role of human papillomavirus in risk management. Rev Gynaecol Practice, v.3, n.4, p.229-233. 2003.

De Araujo Souza, P. S. e Villa, L. L. Genetic susceptibility to infection with human papillomavirus and development of cervical cancer in women in Brazil. Mutat Res, v.544, n.2-3, Nov, p.375-83. 2003.

De Villiers, E. M. Heterogeneity of the human papillomavirus group. J Virol, v.63, n.11, Nov, p.4898-903. 1989.

De Villiers, E. M., Fauquet, C., *et al*. Classification of papillomaviruses. Virology, v.324, n.1, Jun 20, p.17-27. 2004.

Duensing, S., Lee, L. Y., *et al*. The human papillomavirus type 16 E6 and E7 oncoproteins cooperate to induce mitotic defects and genomic instability by uncoupling centrosome duplication from the cell division cycle. Proc Natl Acad Sci U S A, v.97, n.18, Aug 29, p.10002-7. 2000.

Eddy, S. R. What is a hidden Markov model? Nat Biotechnol, v.22, n.10, Oct, p.1315-6. 2004.

_____. A model of the statistical power of comparative genome sequence analysis. PLoS Biol, v.3, n.1, Jan, p.e10. 2005.

Finzer, P., Aguilar-Lemarroy, A., *et al*. The role of human papillomavirus oncoproteins E6 and E7 in apoptosis. Cancer Lett, v.188, n.1-2, Dec 15, p.15-24. 2002.

Franco, E. L., Duarte-Franco, E., *et al*. Cervical cancer: epidemiology, prevention and the role of human papillomavirus infection. Cmaj, v.164, n.7, Apr 3, p.1017-25. 2001.

Franco, E. L., Villa, L. L., *et al*. Epidemiology of acquisition and clearance of cervical human papillomavirus infection in women from a high-risk area for cervical cancer. J Infect Dis, v.180, n.5, Nov, p.1415-23. 1999.

Frazer, I. Vaccines for papillomavirus infection. Virus Res, v.89, n.2, Nov, p.271-4. 2002.

Frazer, I. H. Prevention of cervical cancer through papillomavirus vaccination. Nat Rev Immunol, v.4, n.1, Jan, p.46-54. 2004.

Furey, T. S., Cristianini, N., *et al*. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, v.16, n.10, Oct, p.906-14. 2000.

Greenblatt, R. J. Human papillomaviruses: Diseases, diagnosis, and a possible vaccine. Clin Microbiol Newsletter, v.27, n.18, p.139-145. 2005.

Greenlee, R. T., Hill-Harmon, M. B., *et al*. Cancer statistics, 2001. CA Cancer J Clin, v.51, n.1, Jan-Feb, p.15-36. 2001.

Greenlee, R. T., Murray, T., *et al*. Cancer statistics, 2000. CA Cancer J Clin, v.50, n.1, Jan-Feb, p.7-33. 2000.

Hardison, R. C. Comparative genomics. PLoS Biol, v.1, n.2, Nov, p.E58. 2003.

Henderson, J., Salzberg, S., *et al*. Finding genes in DNA with a Hidden Markov Model. J Comput Biol, v.4, n.2, Summer, p.127-41. 1997.

Hildesheim, A., Schiffman, M., *et al*. Human papillomavirus type 16 variants and risk of cervical cancer. J Natl Cancer Inst, v.93, n.4, Feb 21, p.315-8. 2001.

Hilleman, M. R. Overview of vaccinology with special reference to papillomavirus vaccines. J Clin Virol, v.19, n.1-2, Oct, p.79-90. 2000.

Holowaty, P., Miller, A. B., *et al*. Natural history of dysplasia of the uterine cervix. J Natl Cancer Inst, v.91, n.3, Feb 3, p.252-8. 1999.

Hrushesky, W. J., Sothorn, R. B., *et al*. Sun exposure, sexual behavior and uterine cervical human papilloma virus. Int J Biometeorol, v.50, n.3, Jan, p.167-73. 2006.

Inca. Atlas de mortalidade por câncer no Brasil (1979-1999). Instituto Nacional de Câncer / Ministério da Saúde. Rio de Janeiro, p.395. 2002

_____. Estimativa 2006 - Incidência de Câncer no Brasil. Instituto Nacional de Câncer / Ministério da Saúde. Rio de Janeiro, p.98. 2005

Jemal, A., Murray, T., *et al*. Cancer statistics, 2003. CA Cancer J Clin, v.53, n.1, Jan-Feb, p.5-26. 2003.

_____. Cancer statistics, 2005. CA Cancer J Clin, v.55, n.1, Jan-Feb, p.10-30. 2005.

Jemal, A., Siegel, R., *et al*. Cancer statistics, 2006. CA Cancer J Clin, v.56, n.2, Mar-Apr, p.106-30. 2006.

Jemal, A., Thomas, A., *et al.* Cancer statistics, 2002. CA Cancer J Clin, v.52, n.1, Jan-Feb, p.23-47. 2002.

Jemal, A., Tiwari, R. C., *et al.* Cancer statistics, 2004. CA Cancer J Clin, v.54, n.1, Jan-Feb, p.8-29. 2004.

Jimenez-Sanchez, G., Childs, B., *et al.* Human disease genes. Nature, v.409, n.6822, Feb 15, p.853-5. 2001.

Jorissen, R. N. e Gilson, M. K. Virtual screening of molecular databases using a support vector machine. J Chem Inf Model, v.45, n.3, May-Jun, p.549-61. 2005.

Joung, J.-G., O, S. J., *et al.* Prediction of the Risk Types of Human Papillomaviruses by Support Vector Machines. Lecture Notes in Computer Science, p.723-731. 2004.

Joung, J. G., June, O. S., *et al.* Protein sequence-based risk classification for human papillomaviruses. Comput Biol Med, Aug 2. 2005.

Korf, I. Gene finding in novel genomes. BMC Bioinformatics, v.5, May 14, p.59. 2004.

Kotecha, M. T., Afghan, R. K., *et al.* Enhanced tumour growth after DNA vaccination against human papilloma virus E7 oncoprotein: evidence for tumour-induced immune deviation. Vaccine, v.21, n.19-20, Jun 2, p.2506-15. 2003.

Legood, R., Gray, A., *et al.* Lifetime effects, costs, and cost effectiveness of testing for human papillomavirus to manage low grade cytological abnormalities: results of the NHS pilot studies. Bmj, v.332, n.7533, Jan 14, p.79-85. 2006.

Lenhard, B., Sandelin, A., *et al.* Identification of conserved regulatory elements by comparative genome analysis. J Biol, v.2, n.2, p.13. 2003.

Lesk, A. M. Introduction to Bioinformatics. New York: Oxford University Press Inc. 2002. 255 p.

Leslie, C. S., Eskin, E., *et al.* Mismatch string kernels for discriminative protein classification. Bioinformatics, v.20, n.4, Mar 1, p.467-76. 2004.

Lorenzato, F., Ho, L., *et al.* The use of human papillomavirus typing in detection of cervical neoplasia in Recife (Brazil). Int J Gynecol Cancer, v.10, n.2, Mar, p.143-150. 2000.

Lorenzato, F., Singer, A., *et al.* Cervical cancer detection by hybrid capture and evaluation of local risk factors. Int J Gynaecol Obstet, v.73, n.1, Apr, p.41-6. 2001.

Majoros, W. H., Pertea, M., *et al.* Efficient implementation of a generalized pair hidden Markov model for comparative gene finding. Bioinformatics, v.21, n.9, May 1, p.1782-8. 2005.

Mathe, E., Olivier, M., *et al.* Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. Nucleic Acids Res, v.34, n.5, p.1317-25. 2006.

Matsukura, T. e Sugase, M. Relationships between 80 human papillomavirus genotypes and different grades of cervical intraepithelial neoplasia: association and causality. Virology, v.283, n.1, Apr 25, p.139-47. 2001.

Mccauley, S. e Hein, J. Using hidden Markov models and observed evolution to annotate viral genomes. Bioinformatics, v.22, n.11, Apr 13, p.1308-1316. 2006.

Mcmurray, H. R., Nguyen, D., *et al.* Biology of human papillomaviruses. Int J Exp Pathol, v.82, n.1, Feb, p.15-33. 2001.

Melanie, M. An Introduction to Genetic Algorithms. Cambridge: A Bradford Book The MIT Press. 1999

Modis, Y., Trus, B. L., *et al.* Atomic model of the papillomavirus capsid. Embo J, v.21, n.18, Sep 16, p.4754-62. 2002.

Molijn, A., Kleter, B., *et al.* Molecular diagnosis of human papillomavirus (HPV) infections. J Clin Virol, v.32 Suppl 1, Mar, p.S43-51. 2005.

Motoyama, S., Ladines-Llave, C. A., *et al.* The role of human papilloma virus in the molecular biology of cervical carcinogenesis. Kobe J Med Sci, v.50, n.1-2, Jan, p.9-19. 2004.

Mount, D. W. Bioinformatics: Sequence and Genome Analysis. New York: Cold Spring Harbor Laboratory Press. 2001. 565 p.

Munger, K., Baldwin, A., *et al.* Mechanisms of human papillomavirus-induced oncogenesis. J Virol, v.78, n.21, Nov, p.11451-60. 2004.

Munger, K. e Halpern, A. L. HPV16 E7: Primary Structure and Biological Properties. In: G. Myers, Sverdrup, F., *et al* (Ed.). Human Papillomaviruses 1997 Compendium. Los Alamos: Los Alamos National Laboratory, 1997. p.III (17-36)

Munoz, N. Human papillomavirus and cancer: the epidemiological evidence. J Clin Virol, v.19, n.1-2, Oct, p.1-5. 2000.

Munoz, N., Bosch, F. X., *et al.* Epidemiologic classification of human papillomavirus types associated with cervical cancer. N Engl J Med, v.348, n.6, Feb 6, p.518-27. 2003.

Oh, Y. L., Shin, K. J., *et al.* Significance of high-risk human papillomavirus detection by polymerase chain reaction in primary cervical cancer screening. Cytopathology, v.12, n.2, Apr, p.75-83. 2001.

Parkin, D. M., Bray, F., *et al.* Global cancer statistics, 2002. CA Cancer J Clin, v.55, n.2, Mar-Apr, p.74-108. 2005.

Pedersen, J. S. e Hein, J. Gene finding with a hidden Markov model of genome structure and evolution. Bioinformatics, v.19, n.2, Jan 22, p.219-27. 2003.

Perna, N. T., Plunkett, G., 3rd, *et al.* Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature, v.409, n.6819, Jan 25, p.529-33. 2001.

Prosdocimi, F., Cerqueira, G. C., *et al.* Bioinformática: Manual do Usuário. Biotecnologia Ciência & Desenvolvimento, n.29, p.12-25. 2003.

Ratnam, S., Franco, E. L., *et al.* Human papillomavirus testing for primary screening of cervical cancer precursors. Cancer Epidemiol Biomarkers Prev, v.9, n.9, Sep, p.945-51. 2000.

Slomovitz, B. M., Sun, C. C., *et al.* Are women ready for the HPV vaccine? Gynecol Oncol, Mar 19. 2006.

Souto, M. C. P., Lorena, A. C., *et al.* Técnicas de Aprendizado de Máquinas para Problemas em Biologia Molecular. In: Sbc (Ed.). Anais da III Jornadas de Mini-Cursos de Inteligência Artificial, v.VIII, 2003. p.103-152

Stanley, M. HPV vaccines. Best Pract Res Clin Obstet Gynaecol, v.20, n.2, Apr, p.279-93. 2006.

Steller, M. A. Cervical cancer vaccines: progress and prospects. J Soc Gynecol Investig, v.9, n.5, Sep-Oct, p.254-64. 2002.

Stern, P. L. Immune control of human papillomavirus (HPV) associated anogenital disease and potential for vaccination. J Clin Virol, v.32 Suppl 1, Mar, p.S72-81. 2005.

Swan, D. C., Rajeevan, M., *et al.* Human papillomavirus type 16 E2 and E6/E7 variants. Gynecol Oncol, v.96, n.3, Mar, p.695-700. 2005.

Tristram, A. HPV information needs. Best Pract Res Clin Obstet Gynaecol, v.20, n.2, Apr, p.267-77. 2006.

Van Rossum, A. G., Schuurin-Scholtes, E., *et al.* Comparative genome analysis of cortactin and HS1: the significance of the F-actin binding repeat domain. BMC Genomics, v.6, n.1, p.15. 2005.

Villa, L. L., Sichero, L., *et al*. Molecular variants of human papillomavirus types 16 and 18 preferentially associated with cervical neoplasia. J Gen Virol, v.81, n.Pt 12, Dec, p.2959-68. 2000.

Waack, S., Keller, O., *et al*. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC Bioinformatics, v.7, n.1, Mar 16, p.142. 2006.

Wentzensen, N., Vinokurova, S., *et al*. Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. Cancer Res, v.64, n.11, Jun 1, p.3878-84. 2004.

Whitley, D. A Genetic Algorithm Tutorial Statist Comput, v.5, p.65-85. 1994.

Wu, J. S., Lee, C., *et al*. Primer design using genetic algorithm. Bioinformatics, v.20, n.11, Jul 22, p.1710-7. 2004.

Yamada, T., Manos, M. M., *et al*. Human papillomavirus type 16 sequence variation in cervical cancers: a worldwide perspective. J Virol, v.71, n.3, Mar, p.2463-72. 1997.

Yu, U., Lee, S. H., *et al*. Bioinformatics in the post-genome era. J Biochem Mol Biol, v.37, n.1, Jan 31, p.75-82. 2004.

Zur Hausen, H. Papillomaviruses causing cancer: evasion from host-cell control in early events in carcinogenesis. J Natl Cancer Inst, v.92, n.9, May 3, p.690-8. 2000.

_____. Cervical carcinoma and human papillomavirus: on the road to preventing a major human cancer. J Natl Cancer Inst, v.93, n.4, Feb 21, p.252-3. 2001.

_____. Papillomaviruses and cancer: from basic studies to clinical application. Nat Rev Cancer, v.2, n.5, May, p.342-50. 2002.

**ARTIGO - INFERENCE OF CARCINOGENICITY OF HUMAN
PAPILLOMAVIRUSES BASED ON HIDDEN MARKOV MODEL AND
SUPPORT VECTOR MACHINES**

Artigo enviado a revista
Bioinformatics

As normas para publicação na revista
Bioinformatics encontram-se disponíveis
no site da revista.

http://www.oxfordjournals.org/bioinformatics/for_authors/index.html

Dear Dr.

Raul Fonseca Neto, David A.L. Moraes, Rafael S. Oliveira, Carlos H.M. Castelletti, Saul C. Leite, Rangel C. Souza, Thaís Gaudencio, José L. Lima Filho and Ana T.R. Vasconcelos

Your manuscript entitled:

Inference of Carcinogenicity of Human Papillomaviruses Based on Hidden Markov Model and Support Vector Machines

has been received by Bioinformatics and has been assigned the number: BIOINF-2006-0285

We will contact you when an editorial decision has been made for your manuscript. Meanwhile you can track the progress of your manuscript by looking in the Submitted Manuscripts section of your Author Center in Manuscript Central or by contacting the Editorial Office at Bioinformatics@editorialoffice.co.uk.

Thank you for submitting your paper to Bioinformatics.

Sincerely,
Ms. Abbi Cliff
Administrator, Bioinformatics
<http://bioinformatics.manuscriptcentral.com>

Sequence Analysis

Inference of Carcinogenicity of Human Papillomaviruses Based on Hidden Markov Model and Support Vector Machines

Raul Fonseca Neto^{1,2*}, Dayid A.L. Morais¹, Rafael S. Oliveira², Carlos H.M. Castelletti³, Saul C. Leite¹, Rangel C. Souza¹, Thaís Gaudencio¹, José L. Lima Filho³ and Ana T.R. Vasconcelos¹

¹LNCC Laboratório de Bioinformática, Petrópolis, Rio de Janeiro, Brazil. ²Universidade Federal de Juiz de Fora, Juiz de Fora, Minas Gerais, Brazil. ³Laboratório de Imunopatologia Keizo Asami, Universidade Federal de Pernambuco, Recife, Pernambuco, Brazil

ABSTRACT

Motivation: Prediction of either high-risk or low-risk carcinogenicity of human papillomaviruses (HPV) is an important step towards the understanding of infection mechanisms and treatment of abnormal cells. Currently, some experts predict risk type manually by using molecular procedures or epidemiologic methods. Generative statistical models such as HMM combined with discriminative supervised machine learning approaches such as SVM offer good performance and generalization efficiency in classification tasks that explore biological sequence-based information.

Results: We propose a methodology to classify the carcinogenicity risk type of HPVs by using protein sequences of five genes, namely E6, E1, E7, L1, and L2, which are ubiquitous in 78 HPV genomes. We used a Hidden Markov Model (HMM) to search informative positions in these genes. As a result, a set of twenty-four subsequences was selected using this strategy. Moreover, a discriminative model based on support vector machines was developed using a kernel function that evaluates the similarities among the HPVs based on the chosen subsequences. A genetic algorithm was coupled with the above discriminative model to select a subset of most informative subsequences from this initial set of twenty-four candidates. Consequently, eight subsequences were chosen. Different combinations of these eight regions produced a set of classifiers capable of correctly predicting 100% of HPV risk type in both leave-one-out and ten-fold cross-validation accuracy tests. Three subsequences from gene E6 and one from E7 formed the best set of regions to classify the HPVs. Although no classifier was constructed using the eight selected subsequences at the same time, all subsequences showed an antigenic potential or played an important role in carcinogenicity. The HMM model combined with SVM and a feature selection technique showed to be efficient to distinguish between low- and high-risk HPV, as well as to determine regions that played important roles in HPV carcinogenicity.

Availability: SVM, String Kernel and Genetic Algorithm software were developed at LNCC Bioinformatics Laboratory and are available upon request.

Contact: raulfonsecaneto@ig.com.br

1 INTRODUCTION

Cervical cancer is the third most common cancer among women in the world, after skin and breast cancers. It is responsible for 471,000 new cases and approximately 230,000 deaths per year. In

Brazil, the estimate is 19,260 new cases for 2006, following the same profile of the magnitude observed in the world (Santos et al., 2005).

By the year 2000, new epidemiological evidence, included a large and consistent body of studies, indicating beyond any reasonable doubt, strong and specific associations relating HPV infections to cervical cancer (Bosch, 2002). It has been demonstrated that more than 99% of cervical cancers have detectable human papillomaviruses (HPV) that infect epithelial tissues (Cruickshank, 2003; Wallin et al., 1999) and a common feature of the vast majority is the presence of high risk type (Janicek and Averette, 2001). Due to this association, the World Health Organization (WHO) recognized the first cancer 100% attributable to an infection (Frazer, 2004). A number of 118 HPV genotypes were classified according to their biological niche, oncogenic potential, and phylogenetic position (Bernard, 2005; de Villiers et al., 2004).

Of most importance for diagnosis and therapy is to discriminate which HPV genotypes are highly risky (Joung et al., 2004). There are few methods to test the carcinogenicity of new HPVs isolated from patients. Abnormal cervical epithelial cells can be detected microscopically following Papanicolaou (Pap) staining. Conventional cervical smears or homogeneous cell suspensions from a liquid cytology medium are used for this purpose. Molecular detection of HPV provides a different approach to screening and patient management (Molijn et al., 2005).

In this study, we propose a methodology to classify the carcinogenicity risk of HPVs by using protein sequences of five genes, E6, E1, E7, L1, and L2, which are present in 78 HPV genomes, characterized according to the risk of cervical cancer and availability in the NCBI. We used a generative statistical or Hidden Markov model (HMM) to search for the most informative positions in these genes. Also, a discriminative model based on support vector machine (SVM) was developed using a string kernel function (Fonseca Neto et al., 2004) that evaluates the similarity among the HPVs based on their protein sequences. When a kernel function is used in conjunction with SVMs, the input sequences are implicitly mapped into a high-dimensional vector space. Therefore, we do not need to have an explicit representation of the mapping. It suffices to be able to compute the kernel function for all pairs of virus sequences producing a matrix that represents the degree of similarity between all HPV sequences. A genetic algorithm was coupled with the discriminative model to implement a wrapper feature selection method used to find a subset of the most informative subsequences. The performance of the proposed method was

*To whom correspondence should be addressed.

evaluated by means of a leave-one-out and a ten-fold cross-validation test.

2 DATA SET AND GENERATIVE MODEL

2.1 Data Set Resource

The data on the complete genome of 78 HPV were downloaded from Entrez genome (<http://www.ncbi.nlm.nih.gov>). The classification of the HPV risk type was based on oncogenic potential and phylogenetic position (Bernard, 2005; de Villiers et al., 2004; Munoz et al., 2003). Usually, the characterization of new types of HPV is made based on differences in nucleotide sequence homology compared with other defined HPV types (Pfister et al., 1986). Furthermore, there is no quick method to test if the new HPVs are low or high risk (Joung et al., 2004).

2.2 Pre-Processing with HMM

The five genes, E6, E1, E7, L1, and L2, present in the 78 HPV genomes were selected to be part of the training and test sets. The procedure for selecting the most informative regions in each gene was as follows: Initially, the genes were aligned using the ClustalX program. Subsequently, the aligned sequences were divided into low- and high-risk groups according to their HPV manual classification. The aligned sequences from both groups were subdivided into blocks of 12 amino acids with an overlap of 9 amino acids with the adjacent blocks. Fifteen most representative viruses from both low- and high-risk types were selected based on the literature to construct the HMM models. Two models were built for each aligned subsequences using the HMMER program, one for low-risk and the other for high-risk type. Next, all sequences from the chosen genes were screened with the two HMM and the log-odds scores calculated from the positive (high-risk) and negative (low-risk) models. Finally, the log-odds score difference for each screened region was calculated.

2.3 Selection of Subsequences

After the above procedure, the subsequences with the highest differences were chosen for classification. In order to consider longer sequences, adjacent regions with significant score differences were joined. The final size of these regions varied from 12 to 36 amino acids. As a result, twenty-four subsequences from E6 (four regions), E1 (five regions), E7 (four regions), L1 (six regions) and L2 (five regions) were chosen. This sequence-based information was used in a discriminative analysis that employed a string based kernel SVM classifier with a feature selection algorithm in order to produce a predictive model.

3 DISCRIMINATIVE ANALYSIS AND PREDICTIVE MODEL

3.1 Overview of Support Vector Classifiers

Support vector machine (SVM), first introduced by Boser et al. (1992) and later by Cortes and Vapnik (1995), is an algorithm used to construct binary classifiers that will predict the class of future unseen data. The problem of binary classification may be stated as follows: given the training set $\{(x_i, y_i)\}_{i=1, \dots, m}$, $(x_i, y_i) \in \mathcal{R}^n \times \{-1, +1\}$, construct a function $f: \mathcal{R}^n \rightarrow \{-1, +1\}$, called a classifier, that will “correctly” assign classes y_{m+k} to new points x_{m+k} , $k \in N$. The

SVM is called a maximal margin linear classification algorithm since it searches for linear classifiers of the form:

$$f(x_i) = \text{sign}(\langle w, x_i \rangle + b)$$

where $b \in \mathcal{R}$ is the bias parameter and $w \in \mathcal{R}^n$ is the normal vector of the hyperplane that splits the data, and the term *maximal margin* is attributed to the fact that it constructs a hyperplane that maximizes its minimal distance from points of opposite classes. In order to construct this hyperplane, one needs to solve the following optimization problem for a hard margin SVM:

$$\text{Min} \|w\|_2$$

s.t.

$$y_i(\langle w, x_i \rangle + b) \geq 1, \forall i \in \{1, \dots, m\}.$$

Introducing a set of nonnegative Lagrangian multipliers $\{\alpha_i\}_{i=1, \dots, m}$, the above formulation can be redefined in its dual form. This particular formulation is called Wolfe’s dual:

$$\max L(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

s.t.

$$\sum_i \alpha_i y_i = 0$$

$$\alpha_i \geq 0, \forall i \in \{1, \dots, m\}$$

Solving this problem and finding the optimal α^* values, we can obtain the hyperplane’s normal vector by:

$$w^* = \sum_i \alpha_i^* y_i x_i$$

3.2 Kernel Methods

Kernel methods were first introduced in large margin classifiers as a way of solving nonlinearly separable data (Boser et al., 1992). However, they quickly became a powerful tool for solving non-numerical problems. This way, kernels can be coupled with SVMs to classify sequence-based information.

Formally, we introduce the kernel by the mapping function $\Phi: X \rightarrow F$, which takes points x_i in the input space X and maps them to the so-called feature space F . So we define the kernel function, $k: X \times X \rightarrow \mathcal{R}$, as the inner product of the feature vectors $\Phi(x_i)$ in the feature space F , that is, $k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_F$. However, we do not need to explicitly know the mapping Φ . The kernel function k is constructed directly as a function used to measure similarities between points in input space X .

Having defined the kernel function, we can replace every instance of inner products in $\mathcal{R}^n$ in the SVM dual formulation. Also, given the definition of the normal vector w in the last subsection and the bias value b , we can construct a new discriminant function:

$$f(x_i) = \text{sign}(\sum_j \alpha_j y_j k(x_i, x_j) + b)$$

Note that in order to improve computational efficiency, we can compute the kernel function for each pair of points (x_i, x_j) once and store them in a matrix K , called kernel matrix. Clearly, not every function defined in an input space can be a kernel, and for it to be one, it needs to attend the Mercer conditions (Mercer, 1909).

3.3 String Kernel

In order to construct a kernel for strings, we define the degree of similarity between two biological sequences as the number of matching subsequences, as in (Leslie et al., 2002). However, we allow the length of each subsequence to vary from k_i to k_f . For example, given two sequences x_i and x_j we will compare every subsequence of length $l \in \{k_i, \dots, k_f\}$ for an exact match, and the similarity score will be proportional to the number of matches. For computational efficiency, we use a suffix table to speed up the matching algorithm. For example, considering $k_i=1$ and $k_f=8$, the suffix table for sequence $x_i = (C, T, K, F, C, T, C, K, C, F)$, for an amino acid alphabet, is given by Table 1. The symbol ϵ represents the end of a subsequence and does not count as a matching score. The columns of the table indicate the length of the subsequences that vary from k_i to k_f . The rows of the table are constructed by shifting the sequence to the left until an empty string is found. Therefore, the suffix table implements an easy way to quickly access every subsequence of varying lengths of a given sequence. For example, if we want the second subsequence of size $l=5$, we go to the second row of the table and select the aminoacids from columns 1 through 5. Observe that the suffix table can be constructed in linear time.

Table 1. Suffix table for sequence $x_i = (C, T, K, F, C, T, C, K, C, F)$ and $k_i=1$, $k_f=8$. This table is an efficient representation of the substrings present in a string. Each substring can be referenced indirectly by the integer vector, (9,7,5,1,10,4,8,3,6,2), constructed by sorting the rows of the table in lexicographical order.

	1	2	3	4	5	6	7	8
1	C	T	K	F	C	T	C	K
2	T	K	F	C	T	C	K	C
3	K	F	C	T	C	K	C	F
4	F	C	T	C	K	C	F	ϵ
5	C	T	C	K	C	F	ϵ	
6	T	C	K	C	F	ϵ		
7	C	K	C	F	ϵ			
8	K	C	F	ϵ				
9	C	F	ϵ					
10	F	ϵ						

The matching algorithm, for a fixed substring length l , proceeds by comparing the rows of the suffix table of each sequence, x_i and x_j , starting at column 1 through column l for an exact match. The algorithm returns the number of identical rows. In order to obtain linear complexity for the matching algorithm, it is necessary to sort the rows of the suffix table. This can be made in complexity of the order $O(kn)$, by a recursive bucket sort, where k denotes the length of the substring and n the length of the sequence, and does not depend on the size of the alphabet. After sorting, the suffix table can be represented by an integer vector where the subsequences are accessed indirectly, as summarized in Table 1. It is important to observe that this feature vector representation is linear in space and is more efficient than suffix tree data structure representations (Ukkonen, 1995).

In this way, the kernel function is calculated as follows:

$$k(x_i, x_j) = \left(\frac{1}{(k_f - k_i + 1)} \right) \sum_{l=k_i}^{k_f} \lambda_l M_l(x_i, x_j),$$

where λ_l is a weight function that depends on the size of the subsequence, and $M_l(x_i, x_j)$ is the number of matches on substrings of size l occurring in both x_i and x_j , which is the result of the matching algorithm. Due to the convenience of values in the interval $[0,1]$ we need to divide each result by the maximum value that can be reached. This value is obtained when both sequences are equal, then: $M_l(x_i, x_j) = n - l + 1$, where n is the length of the shortest of the two sequences. Therefore, we define λ_l as:

$$\lambda_l = \frac{1}{(n - l + 1)}, l \in \{k_i, \dots, k_f\}$$

So the final kernel function is given by the following expression:

$$k(x_i, x_j) = \sum_{l=k_i}^{k_f} \frac{M_l(x_i, x_j)}{(n - l + 1)(k_f - k_i + 1)}.$$

Throughout this work several different substring lengths were considered and tested. We opted for $k_i=3$ and $k_f=4$, which produced the best results for SVM classification using the selected regions of genes E6, E1, E7, L1, and L2 individually. This choice is similar to the one made by Joung et al. (2004) for gene E6, where the authors used the mismatch string kernel developed by Leslie et al. (2004), with subsequence length 4 and mismatch 0 and 1.

3.4 Data Fusion

With the kernel function defined as in last section, we created several kernel matrices, one for each selected subsequence. Therefore, in order to compute the total similarity between two HPVs, it was necessary to devise a strategy to fuse this information altogether. We opted for a method called *intermediate integration* (Pavlidis et al., 2002) in conjunction with a feature selection approach, which considers the formation of a new matrix, which we call fusion matrix, by a combinatorial sum of kernel matrices. This contrasts with the method proposed by Lanckriet et al. (2004) that considers a weighted linear combination of kernel matrices solved by semi-definite programming.

In order to choose subsequences that were most influential in classification, several different configurations of the fusion matrix were considered using a genetic algorithm (GA). Each configuration represented a different combinatorial sum of the kernel matrices. This way, each combination was represented by a binary vector in the GA population, where 1 (one) represents the presence of a kernel matrix and 0 (zero) the absence.

To maintain the values of the fusion matrix in the closed interval $[0,1]$, we divided them by the number of kernel matrices present in the sum. Also, considering the unbalance of the training set, we added the fractions ($\#$ of positive examples)/(total $\#$ of examples) = 20/78 and ($\#$ of negative examples)/(total $\#$ of examples) = 58/78 to the main diagonal of the fusion matrix for the positive and negative examples respectively. This was done, as recommended by Saigo (2004), to correct the bias towards one of the classes.

Table 2. Incremental strategy in Genetic Algorithm for Feature Selection. The black squares represent the selected regions at each stage of the GA. The grey squares represent the included regions that were not selected in the current stage of the GA.

Protein	E6				E1					E7				L1						L2				
Region	R ₁	R ₂	R ₃	R ₄	R ₁	R ₂	R ₃	R ₄	R ₅	R ₁	R ₂	R ₃	R ₄	R ₁	R ₂	R ₃	R ₄	R ₅	R ₆	R ₁	R ₂	R ₃	R ₄	R ₅
Stage1																								
Stage2																								
Stage3																								
Stage4																								
Stage5																								
Stage6																								

Also, as suggested by Shawe-Taylor and Cristianini (2004), this will help maintain the kernel matrix semi-definite positive.

3.5 Feature Selection

The genetic algorithm was used in an incremental strategy to create a list of optimal individuals. These individuals, which represent possible subsets of the selected subsequences, can be used to generate a rank of final support vector classifiers. The fitness score used in the GA was calculated multiplying the resulting SVM margin by a classification score, that is, the number of correct predictions during a leave-one-out test.

Initially, all four selected regions of E6, the gene that produced individually the best classification, were utilized as the initial population. After a sufficient number of generations, the regions that were most present in the best individuals were kept in the population. The next step was to add the five regions from E1. Once more, the regions present in the most fit individuals, now formed by a subset of regions from the two genes, were kept in the population in a cumulative way. This procedure was repeated for genes E7, L1, and L2, as reported in Table 2. As a result, a final set of eight regions that were present in the best individuals was chosen. With this reduced set, we used a more exhaustive GA search to create a list of best classifiers produced with different subsets of these regions, as shown in Table 3.

Since the leave-one-out test was executed several times during feature selection, it became necessary to implement an efficient procedure to speed-up this test. For that reason, we opted to keep, during all times, the computed values of the kernel function for the test examples on the last line and column of the fusion matrix. Therefore, we guarantee the construction of the classifier without the influence of the test example by reducing the matrix order by one.

3.6 Validation Methodology

The validation of the classifiers was previously done using a leave-one-out test during the feature selection stage. In this case, the leave-one-out score was computed as one component of the fitness function in the genetic algorithm. However, to obtain a more precise and parsimonious validation of the produced results, we also used a k-fold cross-validation test on the best classifiers, listed in Table 3.

Table 3. Results after applying feature selection with SVM and the leave-one-out test. The numbers in the classifier column represents presence and absence of the eight regions (E6_R1, E6_R2, E6_R3, E6_R4, E1_R1, E7_R2, L1_R6, L2_R5).

Classifier	Fitness score	Accuracy (%)	Margin
01110100	18.913238	100	0.242477
01100100	18.765889	100	0.240588
11110100	18.745505	100	0.240327
11110000	18.711837	100	0.239895
11110110	18.551117	100	0.237835
00100100	18.541003	100	0.237705
11010100	18.429061	100	0.236270
11110111	18.427753	100	0.236253
10100100	18.411258	100	0.236042
01110011	18.407990	100	0.236000
11110011	18.376996	100	0.235603
11100111	18.363913	100	0.235435
11111100	18.351974	100	0.235282

It is important to mention that, in the k-fold cross-validation process, the kernel matrices had to be regenerated for each trial since some examples had to be left out of the training set for each experiment. Thus, we eliminated any form of influence of the test examples during the construction of the classifier. We decided to execute ten experiments and therefore divided the dataset into ten subsets of equal size and ratio of positive and negative examples. In each trial we chose nine subsets to train the classifier which was then used to make the prediction in the subset that was left out. The final error was computed as the average of the partial errors and was associated with the classifier. In this sense, the validation methodology that we used was characterized by the final evaluation of the best classifiers produced by the feature selection stages supported by the HMM information. Therefore, the results and information obtained by the generative HMM model and by the feature selection procedure can be interpreted as the utilization of a priori knowledge in the final construction of the model. We

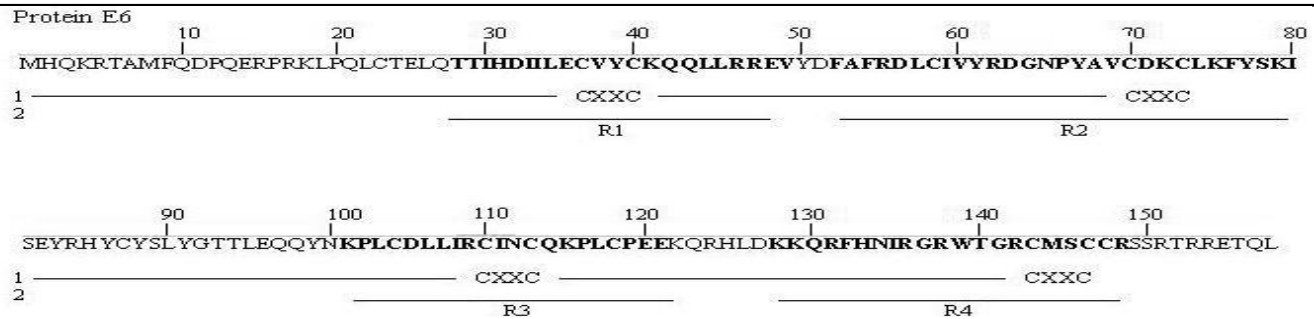


Figure 1. Selected regions of gene E6 after using HMM. Number 1 shows the cysteine that forms the zinc-finger. Number 2 shows the four regions selected for this gene.

understand that the correct choice of the most influential regions is as important as the power of prediction of the classifiers. Therefore, we firstly decided to select the regions and subsequently to make the validation of the classifiers produced by this choice.

4 EXPERIMENTAL RESULTS

4.1 Carcinogenicity

Considering the 78 HPVs used in our dataset, 76 HPVs were classified correctly in all leave-one-out experiments. However, two of them, HPVs 20 and 92, which were initially marked as high-risk type, were classified in all leave-one-out experiments among the low-risk HPVs, appearing in the opposite SVM margin.

Therefore, we changed the label of these two viruses and, contrary to the previous results, the new leave-one-out experiments produced zero error for all best classifiers. This change was not only supported by mathematical and statistical results but also by biological data. According to Antonsson et al. (2003), skin HPV types are widespread among humans, giving rise to sub-clinical infections generally without causing wart or other lesions. In this sense the HPV 20 and 92 are predominantly associated with skin cancer in immunocompromised patients (de Villiers et al., 2004; Forslund et al., 2003) but not for cervical cancers.

4.2 Prediction Performance

Our results, which show 100% empirical prediction for all 78 HPVs in the leave-one-out test, as shown in table 4, were firstly compared with previous works based on Text-Mining approaches. Three methods, namely AdaCost, AdaBoost, and Naive Bayes had been reported in Park et al. (2003). These methods classify the risk type from text data obtained from the HPV Compendiums, which provides a complete list of papillomavirus types and hosts, and shows 93.05, 90.55, and 81.94 % of accuracy (the percentage of answers correctly classified in the dataset) using a leave-one-out test. Later, Joung et al. (2004) developed a classifier that used a string-based kernel SVM and HMM models that shows 93.15 % of accuracy and misclassified five HPV risk types.

Table 4 Predicted and manual classification of HPVs according to their carcinogenicity.

Type	Risk	Type	Risk	Type	Risk
HPV-02a	Low	HPV-28	Low	HPV-58	High
HPV-04	Low	HPV-29	Low	HPV-59	High
HPV-05	Low	HPV-30	High	HPV-60	Low
HPV-06	Low	HPV-31	High	HPV-61	Low
HPV-06a	Low	HPV-32	Low	HPV-62	Low
HPV-06b	Low	HPV-33	High	HPV-63	Low
HPV-07	Low	HPV-34	High	HPV-65	Low
HPV-08	Low	HPV-35	High	HPV-66	High
HPV-09	Low	HPV-36	Low	HPV-67	High
HPV-10	Low	HPV-37	Low	HPV-73	High
HPV-11	Low	HPV-38	Low	HPV-74	Low
HPV-12	Low	HPV-39	High	HPV-75	Low
HPV-13	Low	HPV-40	Low	HPV-76	Low
HPV-15	Low	HPV-43	Low	HPV-77	Low
HPV-16	High	HPV-44	Low	HPV-80	Low
HPV-17	Low	HPV-45	High	HPV-81	Low
HPV-18	High	HPV-47	Low	HPV-82	High
HPV-19	Low	HPV-48	Low	HPV-83	Low
HPV-20	Low*	HPV-49	Low	HPV-84	Low
HPV-21	Low	HPV-50	Low	HPV-85	High
HPV-22	Low	HPV-51	High	HPV-86	Low
HPV-23	Low	HPV-52	High	HPV-87	Low
HPV-24	Low	HPV-54	Low	HPV-89	Low
HPV-25	Low	HPV-55	Low	HPV-90	Low
HPV-26	High	HPV-56	High	HPV-91	Low
HPV-27	Low	HPV-57	Low	HPV-92	Low*

* Initially we used the label high-risk for HPV 20 and HPV 92

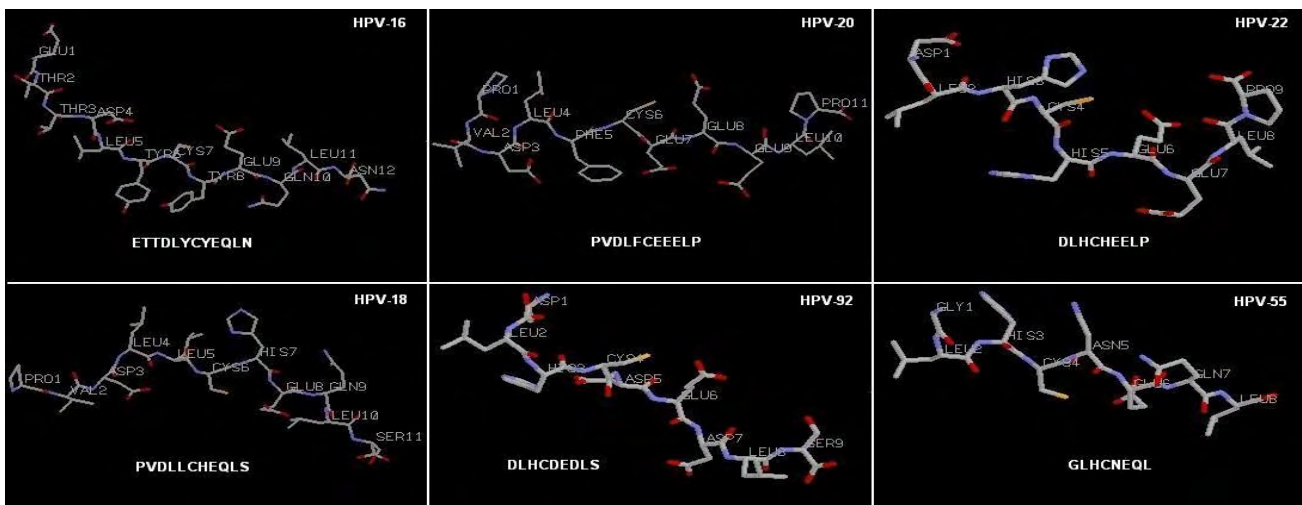


Fig. 2. Binding-site structures of the E7 to the retinoblastoma tumor suppressor. The HPV 16 and HPV 18 are classified as high-risk and the others are classified as low-risk

The greatest difference of the latter method in relation to our approach is that their dataset contained only sequence information about the E6 gene and, therefore, employed neither a data fusion nor a feature selection analysis. Also, it is important to mention that we obtained 100% of empirical accuracy in the ten-fold cross-validation test. These results support that sequence-based classification techniques that combine generative and discriminative data models can show significantly higher classification performance than can text-based classification techniques. These results also show that, for making a 100% correct prediction, sequence information related to more than one gene is needed and, in this case, a data fusion and feature selection approach also has to be employed.

4.3 Population Analysis

Different classifiers constructed from a combination of the eight selected subsequences were capable of correctly classifying the HPV risk types. As we have seen, the most informative regions were four subsequences from E6, one from E1, one from E7, one from L1, and one from L2 (Table 2).

The E6 protein is a basic nuclear and cytoplasmic protein of about 18 kDa (in this study, it will be considered a 158 aa long protein). It contains four cysteine arrays that constitute two relatively large zinc fingers, both of which are required for full function. Either deletion or mutation within the zinc-binding region leads to a structural instability of E6 (Ullman and Emery, 1996). All regions selected in E6 protein have a zinc finger (Figure 1).

The E1 protein is a viral initiator that recognizes the origin of replication (*ori*). The E1 protein binds to the *ori* as a dimer (Enemark et al., 2000). The selected region from E1 was the region responsible for the dimerization. According to Enemark et al. (2000), the dimerization is important for the E1 DNA binding domain (DBD) to bind the *ori* (Figure 3).

The HPV E7 gene encodes a ~21 kDa phosphoprotein. In cervical cancer, the degradation of the retinoblastoma (Rb) tumor suppressor is often through the binding of the E7 protein (Duensing and Munger, 2004; Motoyama et al., 2004). The selected region from E7 is exactly the region responsible for the E7–Rb binding. It

is a nine-amino-acid-long region with a LXCXE motif. A change in some amino acids of this region, especially in this motif, alters the binding capacity of E7 protein (Clements et al., 2000).

To understand the differences between the pRb binding site and the E7 protein from low- and high-risk HPV, we determined the three-dimensional structure of this region (Figure 2). For this, the crystal structure (1GUX) determined by Lee et al. (1998) was used.

Protein E1	
>HPV-16 A	D SIKTLLQQY CLYLHIQ S L A
>HPV-06 S	E AFQKLIEPL SLYAHIQ W L T
Protein L1	
>HPV-16 V	DNR E C I SMDYKQTQLCLIG
>HPV-06 Q	DNR V NVGMDYKQTQLCMVG
Protein L2	
>HPV-16 TV	TT V T THN NP T FTDPSV
>HPV-06 SH	TT T IFR NP V FTEPSV

Figure 3. Shows the regions selected for E1, L1 and L2 proteins. The boxes display the common amino acids between the HPV 16 (high-risk) and HPV 06 (low-risk).

The primary sequence was initially aligned and the three-dimensional structure modeled through the Modeller program. The motif LXCXE was maintained in all high-risk HPV. Some low-risk HPV presented variations of this motif: VXCXN (HPV 10 and 28), LXSXE (HPV 04, 65, 48 and 50), and LXAXE (HPV-60). Some modifications that occur out of the motif and that can change the binding capacity of E7 were also observed.

The HPV encodes two proteins L1 and L2 synthesized late in the infection cycle. The capsid of the virion contains 72 pentamers of L1, centered on the vertices of a T=7 icosahedral lattice (Modis et

al., 2002). L2, a largely internal protein, is present about 1/30 the abundance of L1 and is thought to enhance packaging and infectivity (Roden et al., 1997; Kirnbauer et al., 1993). The L1 selected region is a probable neutralizing epitope (Chen et al., 2000) (Figure 3).

5 DISCUSSION

The HMM and the genetic algorithm proved to be essential for not only generating a good classifier but also for identifying regions that play an important role in the oncogenesis of HPV.

The E6 protein is regarded as a multifunctional protein. It is a transcriptional transactivator, or coactivator. It binds to cellular protein E6AP and E6BP, thereby modulating normal cellular processes. The E6 targets the degradation of p53 and induces the immortalization of primary cell cultures (Duensing and Munger, 2004). All selected regions of E6 participate in one or more activities of this protein. For instance, the four-cysteine arrays indicate two potential zinc fingers with large loops characteristic of a certain class of transcriptional transactivator (Figure 1).

In the E6_R2 a glycine substitution introduced for C70 (cysteine 70, C70G) led to the conclusion that that C70 was essential for *in vitro* zinc binding by HPV 16 E6 protein (Kanda et al., 1991). Crook et al. (1991) predicted the principal E6-p53 binding site to be around 112 to 121. This region may also be important to p53 inactivation (Motoyama et al., 2004) and to accelerate p53 proteasomal degradation (Duensing and Munger, 2004). The p53 can either stimulate or repress gene expression. Pim et al. (1994), analyzing the ability of some mutants of HPV 18 E6 to abolish the p53-mediated transcriptional activation, showed that the wild-type gave 21-fold reduction of this activity, whereas the mutants Δ 94-97, Δ 101-104 and Δ 113-117 (E6_R3, Figure 1), Δ 126-130 and P114G (E6_R4, Figure 1) gave a reduction of 6-fold or less (i.e. to locate these coordinates in HPV 16 E6, add five to the HPV18 E6 coordinate) (Figure 1). Gao et al. (1994) expressed HPV16 E6 peptides in a recombinant vaccinia vector and evaluated the response to these in mice. They found, among others, these two fragments to be antigenic, one of them corresponding to the E6 R3. Park and Androphy (2002) created mutants of HPV E6 proteins. The Y61H and L117Q are defective for p53 degradation (E6 R2 and E6 R4 respectively).

A significant IgA response to the HPV 16 E6 peptide KPLCDLLIRCINCQKPLCPPEE was strongly associated with cervical cancer (Dillner et al., 1995). Notably, this explains the entire region R3 in the Figure 1.

Studies have demonstrated that the replication of HPV depends on the binding of E1 protein at the origin of replication (*ori*) (Enemark et al., 2000). However, to bind the *ori* the E1 needs to be in its dimeric form. The selected region from E1 protein was the region responsible for the dimerization. This may be used as an important target for small molecule intervention of HPV replication.

Although highly conserved, E7 protein of low-risk HPV showed differences that decrease their binding ability. The changes presented in the motif LXCXE by the HPV 10, 28, 04, 65, 48, and 50 severely affect the binding to the retinoblastoma tumor suppressor (Lee et al., 1998). Some low-risk HPV presented the motif LXCXE conserved. However, residues out of the motif have an effect on the binding of E7 protein. The tyrosines at the position three and five in the E7 HPV16 contribute to the stabilization of

the structure (Figure 2). The replacement of the first Tyr with a Phe results in a 10-fold reduction binding activity, and the same substitution of the second Tyr decreases its activity 1000-fold (Jones et al., 1990).

This substitution can be seen in the HPV 20. Similarly, the substitution of Asp, anterior to the motif LXCXE with Asn abolishes the antagonist activity of E7. This Asp, as well as the Glu of the motif, stabilizes the interaction between HPV16 E7 protein and the retinoblastoma tumor suppressor through the formation of a salt bridge (Hu et al., 1990). This can be observed in the HPV55, where the Asp was replaced with Gly. These mutations drastically affect the activity of the E7 protein. Simon et al. (2003) obtained an immune histological reaction in the HPV-infected cervix using a vaccine based on E7 protein.

The L1 protein forms the pentameric assembly unit of the viral capsid. The selected region in our study for L1 protein is similar to that described by Chen et al. (2000) as a possible target to be used in recombinant protein to produce a virus-like particle (VLP). The VLP-based vaccines are effective in preventing papillomavirus infection. According to Harro et al. (2001), an HPV16 L1 vaccine was highly immunogenic and well tolerated, even without adjuvants.

The same L1 region was used to analyze capsomeres of HPV16 for their ability to trigger a cytotoxic T-cell (CTL) response (Ohlschlager et al., 2003). The authors used L1165-173 peptide and observed an induced regression of established tumors expressing L1 determinants (C3 tumor cells).

Chimeras made with L1-L2 showed to be immunogenic and seemed to be a good candidate for a multivalent vaccine (Varsani et al., 2003). The L2 selected region in our study has not been tested as a potential vaccine so far (Figure 3).

6 CONCLUSION

Nowadays, there are discrepancies between epidemiological and phylogenetic classification. The HPV 73 is classified as low risk and high risk by phylogenetic and epidemiological methods, respectively. The epidemiologic classification of HPV 66 in high-risk type was determined by a zero controls and one of three positive cases in biological experiments. In this paper we show a new methodology to classify the HPV risk type, based on mathematical and statistical approach with 100% empirical accuracy. By measuring the similarity of the sequences using both HMM and the kernel function, it was possible to discriminate regions that determined these differences and deduce the type-risk. The two tests, leave-one-out and ten-fold cross validation, supported these results. Additionally, it was shown that even small modifications could be detected by our methodology.

All selected regions found in this paper play an important role for the HPV life cycle. Most of them have been used as a potential vaccine against this virus. 3D models have shown which modifications occurred in the sequences. These modifications suggest that the binding with HPV E7 protein and pRb does not occur.

The identification of the presented regions in these five genes, its biological importance in the viral cycle, and possible connection of these regions with the progress from the lesions to cancer opens possibilities to use this methodology in other illnesses caused by virus, such as dengue and hepatitis.

Clinical tests in women between 25 and 65 years of age reduce the mortality of the cervical cancer by 80%. These regions in high-risk HPV could be used as specific markers to discriminate HPV risk types in biological samples. The identification of a specific papillomavirus type in cervical lesions, warts, and cervix cancer, with our mathematical methodology allied with molecular techniques, allows the development of a method for early and fast diagnosis of the presence of high-risk HPVs and specific treatment for these diseases in the initial phase.

ACKNOWLEDGEMENTS

The authors would like to thank CNPq, “Conselho Nacional de Desenvolvimento Científico e Tecnológico,” for the financial support and the LNCC’s Bioinformatics Laboratory where most of this research took place.

REFERENCES

- Antonsson, A. *et al.* (2003) Prevalence and type spectrum of human papillomaviruses in healthy skin samples collected in three continents. *J Gen Virol*, **84**, 1881-1886.
- Bernard, H.U. (2005) The clinical importance of the nomenclature, evolution and taxonomy of human papillomaviruses. *J Clin Virol*, **32 Suppl 1**, S1-6.
- Boser, B.E. *et al.* (1992) A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*. Pittsburgh, 144-152.
- Bosch, F.X. *et al.* (2002) The causal relation between human papillomavirus and cervical cancer. *J Clin Pathol*, **55**, 244-265.
- Chen, X.S. *et al.* (2000) Structure of small virus-like particles assembled from the L1 protein of human papillomavirus 16. *Mol Cell*, **5**, 557-567.
- Clements, A. *et al.* (2000) Oligomerization properties of the viral oncoproteins adenovirus E1A and human papillomavirus E7 and their complexes with the retinoblastoma protein. *Biochemistry*, **39**, 16033-16045.
- Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Machine Learning*, **20**, 273-297.
- Crook, T. *et al.* (1991) Degradation of p53 can be targeted by HPV E6 sequences distinct from those required for p53 binding and trans-activation. *Cell*, **67**, 547-556.
- Cruikshank, M.E. (2003) The role of human papillomavirus in risk management. *Rev Gynaecol Practice*, **3**, 229-233.
- de Villiers, E.M. *et al.* (2004) Classification of papillomaviruses. *Virology*, **324**, 17-27.
- Dillner, J. *et al.* (1995) Antibodies against linear and conformational epitopes of human papillomavirus type 16 that independently associate with incident cervical cancer. *Int J Cancer*, **60**, 377-382.
- Duensing, S. and Munger, K. (2004) Mechanisms of genomic instability in human cancer: insights from studies with human papillomavirus oncoproteins. *Int J Cancer*, **109**, 157-162.
- Enemark, E.J. *et al.* (2000) Crystal structure of the DNA binding domain of the replication initiation protein E1 from papillomavirus. *Mol Cell*, **6**, 149-158.
- Frazer, I.H. (2004) Prevention of cervical cancer through papillomavirus vaccination. *Nat Rev Immunol*, **4**, 46-54.
- Fonseca Neto, R. *et al.* (2004) A Novel Kernel based methodology to infer similarity among sequences. *Int. Conf on Bioinformatics and Computational Biology*. Rio de Janeiro, Brazil.
- Forslund, O. *et al.* (2003) A broad spectrum of human papillomavirus types is present in the skin of Australian patients with non-melanoma skin cancers and solar keratosis. *Br J Dermatol*, **149**, 64-73.
- Gao, L. *et al.* (1994) Immune response to human papillomavirus type 16 E6 gene in a live vaccinia vector. *J Gen Virol*, **75 (Pt 1)**, 157-164.
- Harro, C.D. *et al.* (2001) Safety and immunogenicity trial in adult volunteers of a human papillomavirus 16 L1 virus-like particle vaccine. *J Natl Cancer Inst*, **93**, 284-292.
- Hu, Q.J. *et al.* (1990) The regions of the retinoblastoma protein needed for binding to adenovirus E1A or SV40 large T antigen are common sites for mutations. *Embo J*, **9**, 1147-1155.
- Janicek, M.F. and Averette, H.E. (2001) Cervical cancer: prevention, diagnosis, and therapeutics. *CA Cancer J Clin*, **51**, 92-114; quiz 115-118.
- Jones, R.E. *et al.* (1990) Identification of HPV-16 E7 peptides that are potent antagonists of E7 binding to the retinoblastoma suppressor protein. *J Biol Chem*, **265**, 12782-12785.
- Joung, J.-G. *et al.* (2004) Prediction of the Risk Types of Human Papillomaviruses by Support Vector Machines. *Lecture Notes in Computer Science*. 723-731.
- Kanda, T. *et al.* (1991) Human papillomavirus type 16 E6 proteins with glycine substitution for cysteine in the metal-binding motif. *Virology*, **185**, 536-543.
- Kimbauer, R. *et al.* (1993) Efficient self-assembly of human papillomavirus type 16 L1 and L1-L2 into virus-like particles. *J Virol*, **67**, 6929-6936.
- Lanckriet, G.R. *et al.* (2004) Kernel-based data fusion and its application to protein function prediction in yeast. *Pac Symp Biocomput*, 300-311.
- Lee, J.O. *et al.* (1998) Structure of the retinoblastoma tumour-suppressor pocket domain bound to a peptide from HPV E7. *Nature*, **391**, 859-865.
- Leslie, C. *et al.* (2002) The spectrum kernel: a string kernel for SVM protein classification. *Pac Symp Biocomput*, 564-575.
- Leslie, C.S. *et al.* (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **20**, 467-476.
- Mercer, J. (1909) Functions of positive and negative type and their connection with theory of integral equations. *Philosophical Transactions of the Royal Society, London A* **209**, 415-446.
- Modis, Y. *et al.* (2002) Atomic model of the papillomavirus capsid. *Embo J*, **21**, 4754-4762.
- Molijn, A. *et al.* (2005) Molecular diagnosis of human papillomavirus (HPV) infections. *J Clin Virol*, **32 Suppl 1**, S43-51.
- Motoyama, S. *et al.* (2004) The role of human papilloma virus in the molecular biology of cervical carcinogenesis. *Kobe J Med Sci*, **50**, 9-19.
- Munoz, N. *et al.* (2003) Epidemiologic classification of human papillomavirus types associated with cervical cancer. *N Engl J Med*, **348**, 518-527.
- Ohlschlager, P. *et al.* (2003) Human papillomavirus type 16 L1 capsomeres induce L1-specific cytotoxic T lymphocytes and tumor regression in C57BL/6 mice. *J Virol*, **77**, 4635-4645.
- Park, R.B. and Androphy, E.J. (2002) Genetic analysis of high-risk e6 in episomal maintenance of human papillomavirus genomes in primary human keratinocytes. *J Virol*, **76**, 11359-11364.
- Park, S.-B. *et al.* (2003) Mining the Risk Types of Human Papillomavirus (HPV) by AdaCost. *Lecture Notes in Computer Science*. 403-412.
- Pavlidis, P. *et al.* (2002) Learning gene functional classifications from multiple data types. *J Comput Biol*, **9**, 401-411.
- Pfister, H. *et al.* (1986) Classification of the papillomaviruses--mapping the genome. *Ciba Found Symp*, **120**, 3-22.
- Pim, D. *et al.* (1994) Mutational analysis of HPV-18 E6 identifies domains required for p53 degradation in vitro, abolition of p53 transactivation in vivo and immortalisation of primary BMK cells. *Oncogene*, **9**, 1869-1876.
- Roden, R.B. *et al.* (1997) Papillomavirus is resistant to desiccation. *J Infect Dis*, **176**, 1076-1079.
- Saigo, H. *et al.* (2004) Protein homology detection using string alignment kernels. *Bioinformatics*, **20**, 1682-1689.
- Santos, A.M.R.d. *et al.* (2005) Estimativa 2006 - Incidência de Câncer no Brasil. INCA/MS, Rio de Janeiro.
- Shawe-Taylor, J. and Cristianini, N. (2004) *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge.
- Simon, P. *et al.* (2003) Cervical response to vaccination against HPV16 E7 in case of severe dysplasia. *Eur J Obstet Gynecol Reprod Biol*, **109**, 219-223.
- Ukkonen, E. (1995) On-line construction of suffix trees. *Algorithmica*, **14**, 249-260.
- Ullman, C.G. and Emery, V.C. (1996) Transforming Proteins of Human Papillomaviruses. *Rev Med Virol*, **6**, 39-55.
- Varsani, A. *et al.* (2003) Chimeric human papillomavirus type 16 (HPV-16) L1 particles presenting the common neutralizing epitope for the L2 minor capsid protein of HPV-6 and HPV-16. *J Virol*, **77**, 8386-8393.
- Wallin, K.L. *et al.* (1999) Type-specific persistence of human papillomavirus DNA before the development of invasive cervical cancer. *N Engl J Med*, **341**, 1633-1638.

CONCLUSÕES E PERSPECTIVAS

O uso da genômica comparativa dos genomas dos tipos de HPV, aliada ao HMM e ao algoritmo genético foi essencial na classificação dos tipos de HPV entre alto e baixo risco. Além disso, com estas metodologias foi possível identificar regiões que podem ser importantes na oncogenicidade do HPV.

Individualmente nenhuma das cinco proteínas estudadas E6, E1, E7, L1, L2, possuía regiões que classificasse os tipos de HPV entre alto e baixo risco.

Diferentes combinações das oito regiões identificadas, E6_R1, E6_R2, E6_R3, E6_R4, E1_R1, E7_R2, L1_R6, L2_R5, podem classificar com 100% de eficiência os tipos de HPV. Entre estas combinações a que obteve a melhor classificação foi a que possui três regiões da E6 e uma da E7.

Estas oito regiões já foram descritas como de importância para a manutenção do vírus, estrutura molecular e sua interação com as proteínas humanas.

As quatro regiões identificadas da proteína E6 possuem resíduos de cisteína e possíveis dedos de zinco que estão associados à ligação com a proteína p53. Mudanças nestas regiões podem acarretar na não ligação da E6 com a p53. A região selecionada da E1 é responsável pela dimerização da proteína e ligação com a origem de replicação. A região E7 encontrada possui o motivo LXCXE, bastante estudado pela sua associação com a habilidade de ligação com a pRb. Mudanças nestas regiões demonstram uma menor capacidade de se ligar a esta proteína como acontece com os HPVs de baixo risco. As regiões das proteínas tardias L1 e L2 são possíveis alvos para o uso em proteínas recombinantes para a produção de partículas semelhantes a vírus (HPV).

Nossa metodologia, utilizando-se das seqüências dos aminoácidos e baseada em uma abordagem matemática e estatística, pode ser um novo método na determinação da classificação de novos tipos de HPV e na confirmação dos já existentes.

Como mencionado anteriormente, existem discrepâncias entre a classificação epidemiológica e filogenética. O modelo apresentado, que se utiliza de seqüências de aminoácidos e abordagem matemática, como na classificação filogenética, corroborou a classificação epidemiológica nas duas divergências encontradas.

A identificação das regiões presentes nestas cinco proteínas, e sua importância biológica no ciclo viral, e as possíveis conexões com o progresso das lesões ao câncer abrem possibilidades para usar esta metodologia em outras doenças causadas por vírus, tal como a dengue e a hepatite.

Estas regiões ainda serão analisadas por outros métodos, não matemáticos, para confirmar sua eficiência na classificação dos tipos de HPV entre alto e baixo risco para o câncer do colo do útero.

Além disso, nosso grupo de HPV do Lika irá desenvolver primers e posterior clonagem destas regiões para um possível diagnóstico de HPV baseado em regiões diferentes dos testes encontrados no mercado.

Com este diagnóstico diferenciado, associado a uma política pública mais eficiente no uso do teste de Papanicolau, pode-se tentar reduzir os custos na detecção do DNA do HPV e aumentar sua utilização como teste secundário e até mesmo primário para os portadores de HPV e do câncer do colo do útero.

ANEXO – ATIVIDADES RELEVANTES DESENVOLVIDAS DURANTE O MESTRADO

Artigos Publicados

- VASCONCELOS, Ana Tereza Ribeiro de; GUIMARÃES, Ana Carolina Ramos; CASTELLETTI, Carlos Henrique Madeiros; *et al.* MamMiBase: a mitochondrial genome database for mammalian phylogenetic studies. *Bioinformatics* (Oxford, England), doi:10.1093, v. 21, n. 10, p. 2566-2567, 2005. Fator de Impacto: 5,742

Cursos

- Seqüenciamento - MegaBACE 750. (Carga horária: 20h). Laboratório de Imunopatologia Keizo Asami, LIKA/UFPE, Brasil, 2006.
- Aprender a Empreender. (Carga horária: 16h). Serviço Brasileiro de Apoio Às Micro e Pequenas Empresas, SEBRAE, Brasil. 2006
- Cristalografia de Proteínas. (Carga horária: 32h). Universidade Federal do Ceará, UFC, Ceará, Brasil. 2005.
- Coleta de Dados na Linha MX1 - LNLS. (Carga horária: 16h). Laboratório Nacional de Luz Síncrotron, LNLS, São Paulo, Brasil. 2005.

- Rede Nacional de Proteômica - Proteômica Nordeste. (Carga horária: 40h). Laboratório de Imunopatologia Keizo Asami, LIKA, Pernambuco, Brasil. 2004.

Disciplinas Ministradas (sob orientação)

- Ciência da Informação. (Carga horária: 45h). Doutorado em Biologia de Fungos. 2005.

Resumos em Congressos e Jornadas

- MORAIS, David Anderson de Lima; F NETO, Raul; OLIVEIRA, Rafael S; CASTELLETTI, Carlos Henrique Madeiros; SOUZA, Rangel C; GAUDENCIO, Thais; LEITE, Saul; LIMA FILHO, José Luiz de; VASCONCELOS, Ana Tereza Ribeiro de. Inference of Carcinogenicity of Human Papilomaviruses Based on Hidden Markov Model and Support Vector Machines. In: X-MEETING 1ST INTERNATIONAL CONFERENCE OF AB3C, 2005, Caxambu. 2005.
- SOARES, Rebecca; CASTELLETTI, Carlos Henrique Madeiros; OLIVEIRA, João Ricardo Mendes de. Definindo o limiar de manifestação dos primeiros sintomas do depósito de Cálcio em Gânglios da Base ("Doença de Fahr") a partir de análise quantitativa de neuroimagem. In: I JORNADA CIENTÍFICA DO LIKA, 2005, Recife. 2005.

- CASTELLETTI, Carlos Henrique Madeiros; LIMA FILHO, José Luiz de. Lactato desidrogenase: Caracterização estrutural e busca de novos cofatores. In: I JORNADA CIENTÍFICA DO LIKA, 2005, Recife. 2005.
- CASTELLETTI, Carlos Henrique Madeiros; GUIMARÃES, Ana Carolina Ramos; CARUSO, *et al.* MAMMIBASE: Mammalian mitochondrial genome database. In: ICOBICOB I - INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND COMPUTACIONAL BIOLOGY, 2004, Angra dos Reis. 2004.

Participação em Eventos

- Icobicobi - International Conference on Bioinformatics and Computacional Biology. Angra dos Reis, 2004.
- I Jornada Científica do LIKA. Recife, 2005.
- Workshop sobre Propriedade Intelectual. Recife, 2006.
- Workshop de Genômica e Biologia Estrutural para Aplicações Biotecnológicas e Médicas. Campinas, 2006.
- Workshop Doenças do Trato Intrauterino. Recife, 2006.