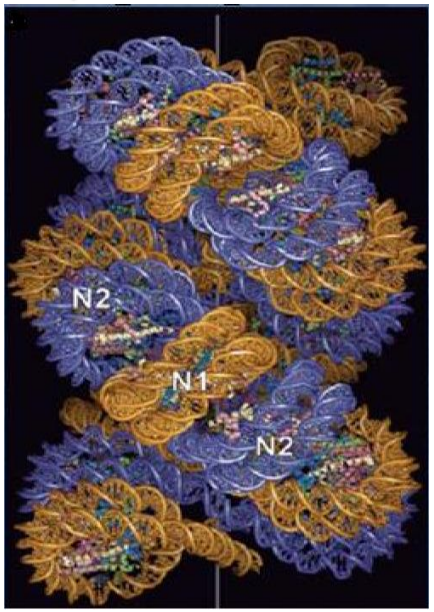
INTERNATIONAL HEREDITARY CANCER CENTER
POMERANIAN MEDICAL UNIVERSITY, SZCZECIN, POLAND

KIEROWNIK: PROFESOR JAN LUBIŃSKI



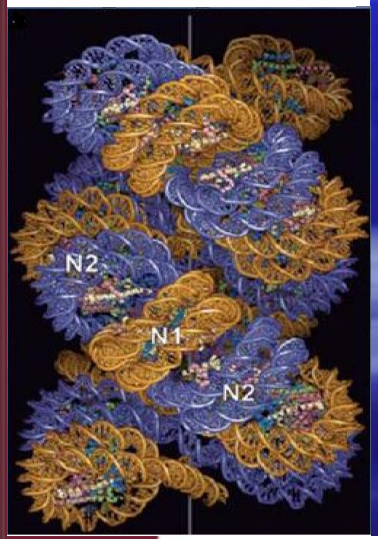
WIESŁAW PIESIAK

Zastosowanie technologii komputerowych do oceny PENETRACJI i RYZYKA zachorowania na RAKA sutka na przykładzie nosicieli mutacji CHEK2 oraz przydatność BIOINFORMATYKI

Konferencję Użytkowników Komputerów Dużej Mocy - Akademickie Centrum Komputerowe Cyfronet

AGH KRAKÓW - ZAKOPANE 12-13.03.2009

GENEZA, FILOZOFIA, PRZEDMIOT, POCHODZENIE I ZAKRES DYSCUPLIN NAUKOWYCH DLA ZDROWIA



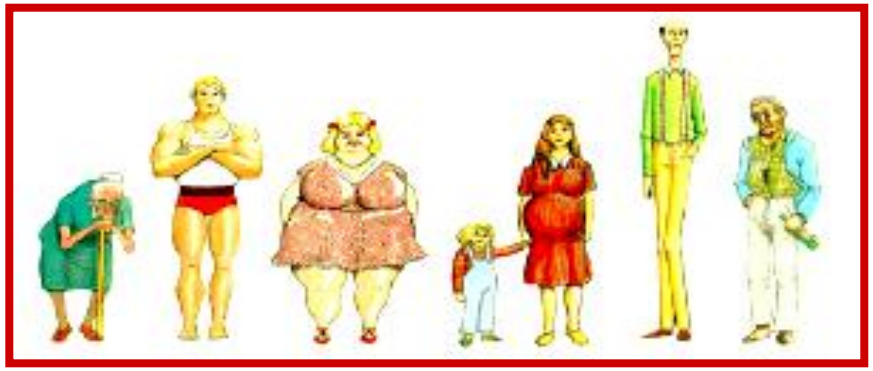
Rys. 1 NOWOTWOTY NADAL SĄ CHOROBĄ ŚMIERTELNĄ. JEST TO O WIELE ZA PÓŹNO NA ETAPIE OGRANICZANIA SKUTKÓW CHOROBY NOWOTWOROWEJ BEZ ANALIZ BIOINFORMATYCZNYCH

CZYNNIKI WPŁYWAJĄCE NA STAN ZDROWIA

1. płeć, ciąża, dieta, wiek, czynniki środowiskowe, genotyp, choroby, masa ciała, interakcje (leki, składniki diety), wchłanianie, dystrybucja, wydalanie, transport, metabolizm

2. Empiryczne strategie farmakoterapii -indywidualna zmienność na podstawie badań klinicznych i analiz BIONFORMATYCZNYCH

Leczenie wszystkich chorych z tym samym rozpoznaniem, tą samą dawką leku



Wiek, płeć, masa ciała, choroby nerek, choroby wątroby, dieta, używki, interakcje, genotyp

Przyczyny zmienionej reakcji na leki –
odmienności farmakogenetyczne -
FARMAKOGENTYKA (pharmacogenetics)

Zmieniona farmakokinetyka

- **Wchłanianie**
- **Dystrybucja**
- **Metabolizm**
- **Wydalanie**

Zmieniona farmakodynamika

- **Receptory**
- **Kanały jonowe**
- **System immunologiczny**
- **Enzymy**

Farmakogenetyka - **Dział farmakologii klinicznej zajmujący się badaniem wpływu genotypu i fenotypu człowieka na działanie i losy leków w organizmie**

Trends in Pharmacol. Sci. 2001,6,298

NIZP- PZH

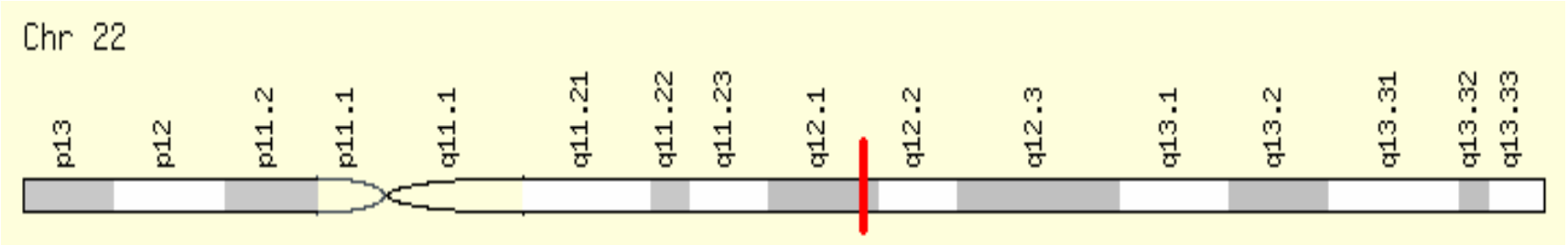
THE ANALITIC LINE OF FAMILIAL AGGR.BREAST, COLON CA AND OTHERS CANCERS- CFA GENES 2000-2009 – CANCEROLOGY

EUROPE PROJECT

EUROPE PROJECT PH - GEN National Task Forces – POLISH NATIONAL TASK FORCES –NIZP - PZH

GEN CHEK2

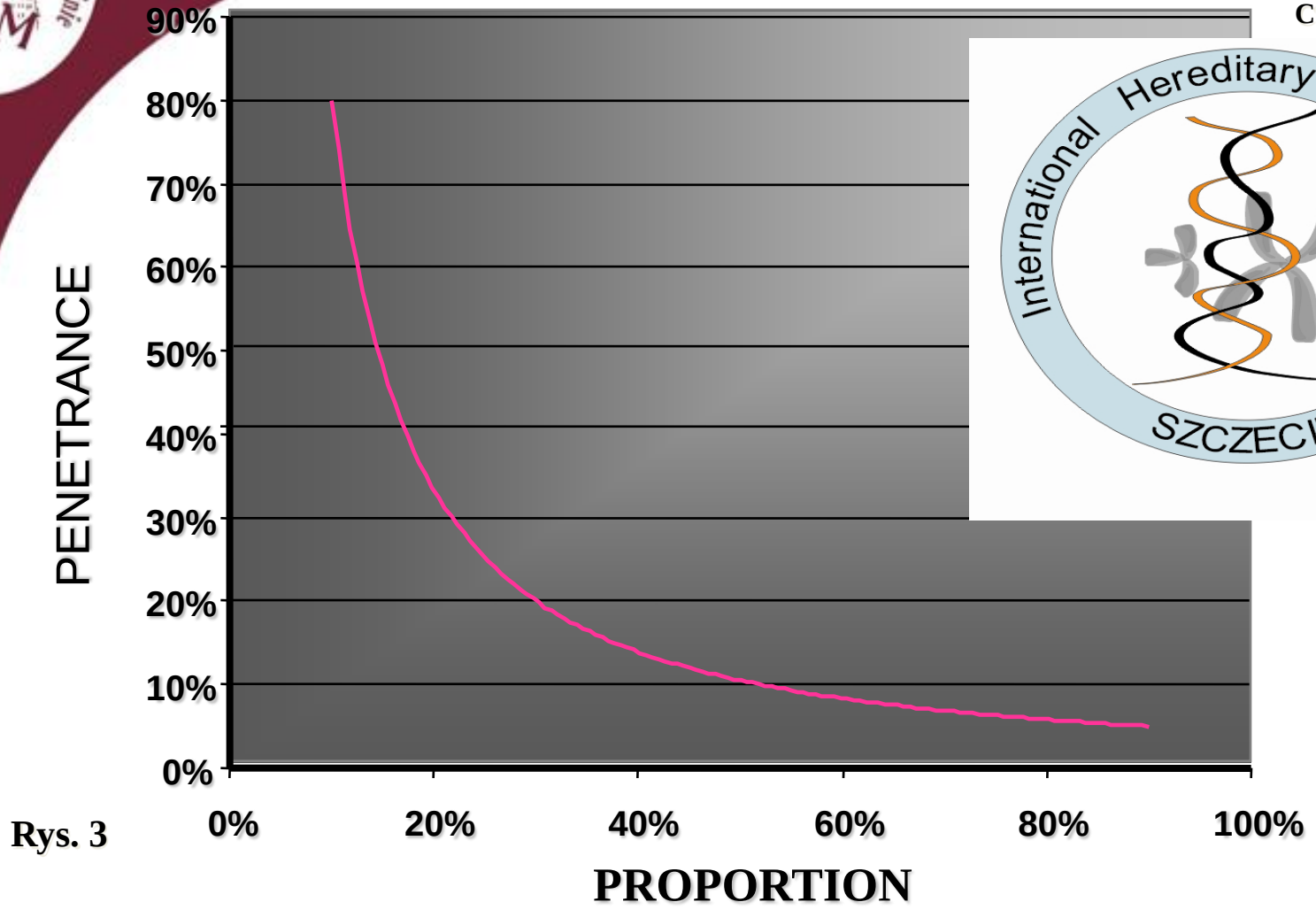
Gene description: checkpoint kinase 2
Maps: **22q12.1**
DNA size: **54,092** kb, 16 exons
Gene type: a protein kinase that is activated in response to DNA damage, is involved in cell cycle arrest



Rys. 2

OBRAZ GRAFICZNY- Lokalizacja genu CHEK2- ten gen **jest antyonkogenem**

PENETRANCE -AND PROPORTION OF CANCERS

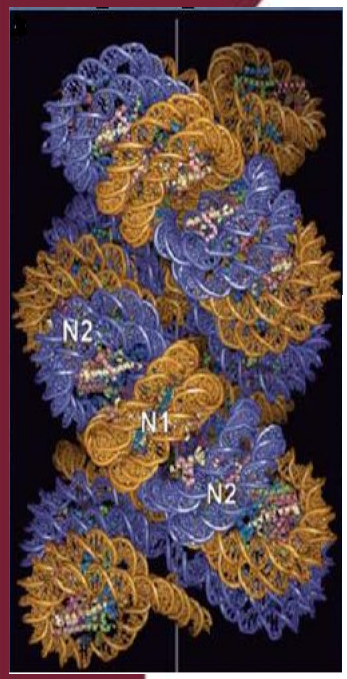


Rys. 3

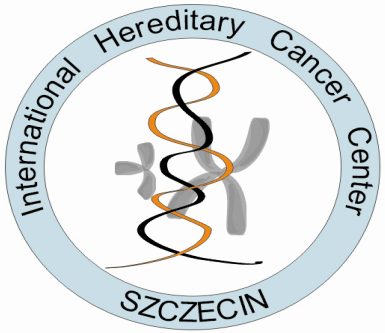
Odsetek zachorowań w ciągu życia na nowotwór w grupie nosicielek mutacji

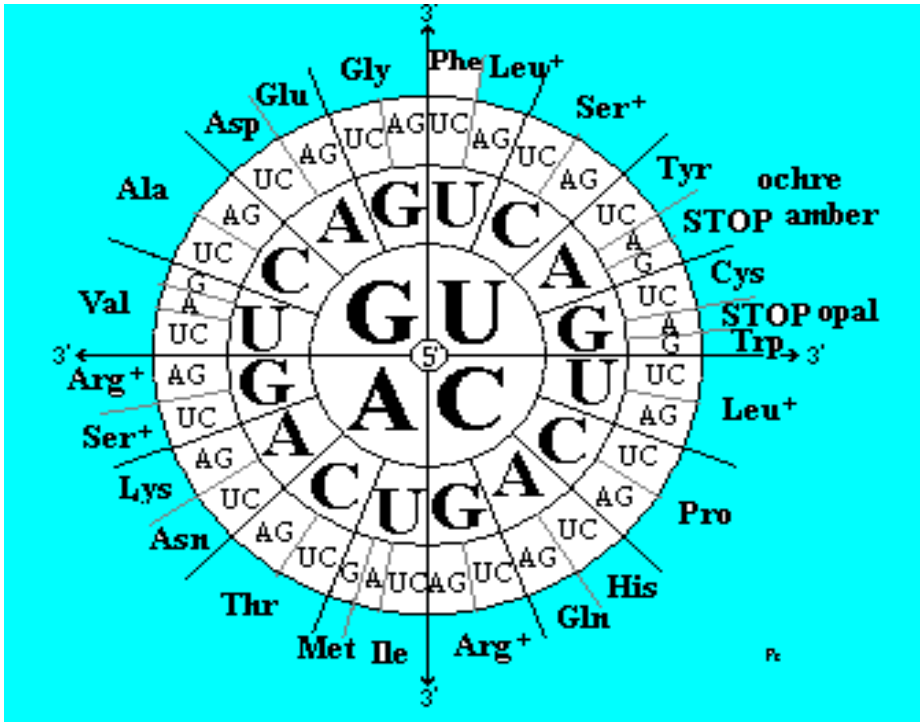
Profesor Lubiński Jan 6.05.2004 Madrid-ESO

GEN EXPRESSION - AND DEFINITION PROBLEM



Ekspresja genu – jest to odczytywanie informacji genetycznej zakodowanej w genie na drodze transkrypcji oraz -w przypadku genów kodujących białka- synteza białka kodowanego przez gen na drodze translacji. Za jeden z podstawowych procesów regulacji ekspresji genów uważa się na przykład: acetylację histonu.





AUTORSKA METODA LICZENIA PENETRACJI I RYZYKA RAKA

1. **Cyrillic2 –** program służący do graficznego sporządzania rodowodów rodzinnych i przechowywania ich **(cf.** http://www.cyrillic.software.com**),**
2. **Program autorski** napisany w języku Java służący ***parsowania*** plików programu Cyrillic2 i przenoszenia danych pomiędzy Cyrillic'iem2 a programami służącymi do dalszych badań,
3. **MATLAB-7.0 –** system pozwalający na programowanie i rozwiązywanie różnych problemów obliczeniowych na cele badań nad w/w zbiorem opracowano w Matlabie zestaw skryptów autorskich, umożliwiających analizy zaawansowane, (cf. http://www.mathworks.com).
4. **WEKA –**otwarty i darmowy system opracowany na Uniwersytecie Waikato napisany w języku Java zawierający duży zestaw gotowych algorytmów z zakresu analizy danych, uczenia maszynowego i tzw .*data- miningu*. Algorytmy te realizują zadania klasyfikacji, aproksymacji, analizy skupień, wykrywania reguł asocjacyjnych w badanym zbiorze danych.(cf. http://www,cs.waikoto.ac.nz/ml/weka/)
5. **Platforma -R –** otwarte i darmowe środowisko przeznaczone do wspomagania obliczeń statystycznych, w szczególności także do przeprowadzania testów statystycznych **(cf: http://www.r-project.org/).**
6. Stosując **metodę Kaplan - Maier- test** przeżycia dla mężczyzn i kobiet - dla krewnych pierwszego rzędu dla liczenia Penetracji oraz **modelu Coxa** dla oceny standaryzowanego przeżycia

Constitutional mutations in the gene CHEK2. [OMIN- Online Mendelian Inheritance in Man-(www.ncbi.nlm.nih.gov/Omim)]

Mutations	OMIM	Phenotype
CHEK2-Ex10 1100 delacja C (Truncating mutations)	604373.0001	Sydrome Li i Fraumeni type 2; predisposition: to breast cancer ; prostate cancer; colon cancer
CHEK2-missense-Ex3-I157T	604373.0002	Sydrome Li i Fraumeni type 2; predisposition: to breast cancer ; prostate cancer; and thyroid cancer, kidney cancer; to many other types of malignant tumors and cancers
p.Arg145Trp	604373.0003	Sydrome Li i Fraumeni type 2
c.1422delT	604373.0004	Sydrome Li i Fraumeni type 2
CHEK2-Ex9/10 del.ecja 5395 (Truncating mutations)	604373.0012	Predisposition to prostate cancer and other types of malignant tumors and cancers
CHEK2- IVS2+1G/T-splice (Truncating mutations)	604373.0013	Predisposition to many types of malignant tumors and cancers : breast and prostate cancers, and thyroid

Tab. nr-1 opisano 18 mutacji w tym genie, w naszym zakładzie badamy-4 CHEK2(niebieski)

CEL PRACY

Wykonanie głównie analizy PENETRACJI mutacji CHEK2 przy zastosowaniu autorskiego programu komputerowego u **2000** kobiet, kolejnych i nie selekcjonowanych zachorowań na raka sutka u krewnych pierwszego stopnia / rodzice, dzieci, rodzeństwo/ oraz określenie RYZYKA zachorowania na raka uwzględniając cztery typy mutacji CHEK2:/-Test CHEK2 dotyczył niżej wymienionych mutacji:

- 1110 delecja A,
- IVS2+1G>A,
- delecja 5395,

Truncating mutations

- I157T----------

Missense mutation

METODY I SPOSOBY LICZENIA PENETRACJI - Statistical analysis

1. ANALIZY (haploview software): wg programu firmy -**SAS-STAT 9.1, SPSS – Statistics**, program firmy **StataCorp LT- STATA 9.0**, firmy StatSoft Polska – **STATYSTYKA-8.0** lub **STATYSTYKA-Data Miner**-stosując dla krewnych pierwszego rzędu metodę **Kaplan – Maier - test** przeżycia dla mężczyzn i kobiet albo też : Metody-**typu *Kin -Cohort Study*** *(* pakiet oprogramowania *MATLAB* –analizy krewnych *kohorty danych* ,zastosowanie **platformy- R** i metody Kaplan - Meier a także **modelu Coxa** dla oceny standaryzowanego przeżycia).

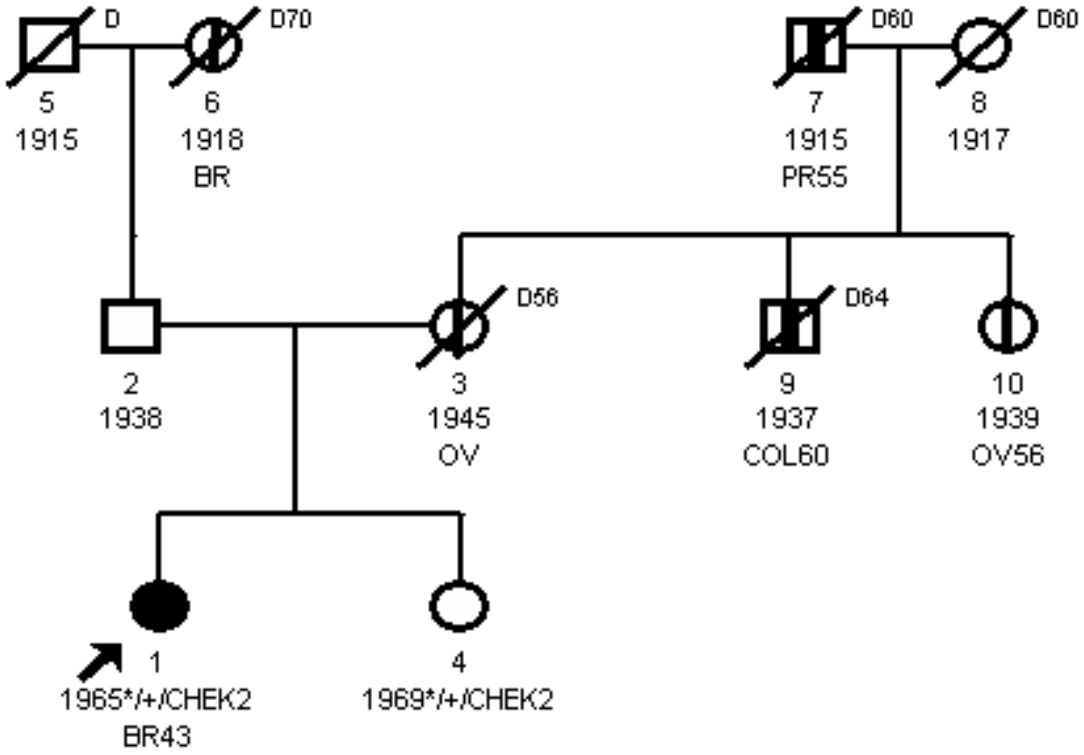
2. Nasza procedura **wyklucza dublowanie czynności** a umożliwia w analizie ocenę statystyczną wszystkich relacji występujących w rodowodzie o ile takie zostaną zgromadzone podczas wykonywania rodowodu. Zamiast korzystać z naszej procedury przekształcania rodzinnego rodowodu z postaci graficznej do zapisu w postaci tabeli dla wykonywanych analiz statystycznych, można wykonać **tabelowy zapis** krewnych pierwszego rzędu z uwzględnieniem wszystkich potrzebnych elementów i następnie dokonać analizy statystyczne stosując odpowiednie statystyki lub procedury analityczne opisane wyżej.

3. Wybór **należy do badaczy** oraz jest zależny od stopnia zaawansowania informatyczno-analitycznego a także umiejętności posługiwania się procedurami informatyczno-statystycznymi przy analizowaniu zgromadzonego materiału badawczego jak również jakie wyniki badawcze oczekują w swoich analizach. Wszystkie metody są bardzo przydatne i z powodzeniem mogą być stosowane.

METODYKA

1. Wykonano testy DNA u 2000 kolejnych nie selekcjonowanych kobiet, które zachorowały na raka sutka.
2. Pobrano również krew do analizy markerów genetycznych DNA od 931 kolejnych pacjentek – grupa kontrolna, które jeszcze nie zachorowały na raka sutka- w grupie tej stwierdzono 61 mutacji CHEK2 u badanych pacjentek.
3. W grupie chorych z mutację CHEK2(jedną z czterech) stwierdzono u 201 pacjentek, krewnych pierwszego stopnia w 186 rodzinach, których rodowody poddano analizie statystycznej stosując autorski programu komputerowego oraz metodę Kaplana - Maiera.
4. W 186 rodzinach w których wystąpiła mutacja CHEK2 uzyskano krew od 240 osób. U 39 pacjentek nie stwierdzono mutacji CHEK2 oraz u 14 kobiet, które zachorowały na raka sutka nie stwierdzono mutacji CHEK2.
5. Jednocześnie mutacja CHEK2 i zachorowanie na raka sutka wystąpiło u 187 pacjentek

PRZYKŁADOWY RODOWÓD RODZINY Z MUTACJĄ



Rys.4 PACJENT Z MUTACJĄ CHEK2 w formie graficznej

RODOWÓD RODZINY Z MUTACJĄ -forma tekstowa

Family ID	ID	Accession no.	Number	Father	Mother	Forename	Surname	Status
	D.O.B.	D.O.D.	Age	Sex	Samp#le	Add inf1	Add inf2	
114755		4	-:4	3	4		IWONA	
				F		1969*/+/CHEK2		
114755		1	-:1	3	4		IWONA	Affected
				F		1965*/+/CHEK2		BR43
114755		2	-:2	5	6		IWONA	
				M		1938		
114755		3	-:3	7	8			Affected
(Hearsay)		?	D56	F		1945	OV	
114755		5	-:5	0	0		IWONA	
		?	D	M		1915		
114755		6	-:6	0	0			Affected
(Hearsay)		?	D70	F		1918	BR	
114755		7	-:7	0	0			Affected
(Hearsay)		?	D60	M		1915	PR55	
114755		8	-:8	0	0		IWONA	
		?	D60	F		1917		
114755		9	-:9	7	8			Affected
(Hearsay)		?	D64	M		1937	COL60	
114755		10	-:10	7	8			Affected
(Hearsay)				F		1939	OV56	

PACJENT Z MUTACJĄ CHEK2 po przekształceniu z postaci graficznej do tekstowej

DATABASE for the calculation of the CHEK2–PENETRANCE or other mutations all unselected cancers cases- the traditional method

PatientID	Today's date	agedx	current age	Gender	deadalive	ethnic	agedead	YOB	bilca	bilagedx	Otca
111231-1		63	0			CHEK2-EX3 I154Tmiss	0	1940		0	Breast
111260-1		48	0			CHEK2-EX3 I154Tmiss	0	1955		0	Breast
111264-1		71	0			CHEK2-EX3 I154Tmiss	0	1932		0	Breast
111329-1		55	0			CHEK2-EX2 IVS2splice	0	1947		0	Breast
111333-1		70	0			CHEK2-EX3 I154Tmiss	0	1933		0	Breast
111339-1		45	0			CHEK2 1100del C	0	1958		0	Breast
111377-1		68	0			CHEK2-EX3 I154Tmiss	0	1935		0	Breast

December-2008. Modified from Prof. Steven Narod database by Wiesław Piesiak and Prof. Jan Lubiński (the cooperative team). For cancer risks in first-degree relatives of CHEK2 mutation carriers: effects of mutation type and cancer site in proband

PLIKI PO PARSOWANIU do analiz statystycznych

family_id, number, EXP16, sex, status, CHEK2HM, tree_level, add_inf_2, CHEK2_father, CHEK2_mother, CHEK2, NOD2, B2P1, EX1P27, CHEK2EX10 CYP1B1, EX1P27_father, EX1P27_mother, EXP16_father

110079, 1, no, F, _, no, 3, _, ?, ?, yes, no, no, no, no, no, ?, ?, ?

110079, 27, no, F, Affected, no, 3, LIV54, ?, ?, no, no, no, no, no, no, ?, ?, ?

110079, 29, no, F, Affected, no, 3, BR57, ?, ?, no, no, no, no, no, no, ?, ?, ?

110181, 1, no, F, Affected, no, 3, SKI-6, ?, ?, yes, no, no, no, no, no, ?, ?, ?

110252, 6, no, F, _, no, 4, _, ?, yes, yes, no, no, yes, no, no, ?, yes, ?

110252, 1, no, F, Affected, no, 3, SKI47, ?, ?, yes, no, no, yes, no, no, ?, ?, ?

110289, 27, no, F, _, no, 4, _, ?, yes, no, no, no, no, no, no, ?, no, ?

110289, 1, no, F, Affected, no, 3, BR50, ?, ?, yes, no, no, no, no, no, ?, ?, ?

110289, 28, no, F, _, no, 4, _, ?, yes, no, no, no, no, no, no, ?, no, ?

110336, 4, no, F, _, no, 3, _, ?, ?, yes, no, no, yes, no, no, ?, ?, ?

110336, 31, no, F, _, no, 3, _, ?, ?, yes, no, no, no, no, no, ?, ?, ?

TEORIA-OPIS PROCEDURY ANALITYCZNO-INFORMATYCZNEJ

1. Zapisane w programie **Cyrillic-2** poszczególne rodowody rodzinne były zapisywane z postaci binarnej-graficznej (rozszerzenie .*fam*– format programu Cyrillic-2) do postaci plików tekstowych – jest to opcją samego programu Cyrillic-2

2. Następnie zbiór plików tekstowych poddawany był **parsowaniu** za pomocą napisanego przez autora i informatyków programu w języku Java.

3. Po sparsowaniu **wszystkich** plików, zbiór przekształcany był do jednego już tylko pliku tekstowego.

4. Można powiedzieć, że na tym etapie zbiór rodowodów został przekształcony do postaci typowej tabeli danych, w której wierszami pisani są pacjenci (lub członkowie rodzin) a kolumnami zmienne ich opisujące. Pomimo przekształcenia rodowodów do tabeli, kolumny w niej zawarte zachowują pełną informację zawartą w oryginalnych rodowodach.

5. Uzyskany w ten sposób plik mógł być już dalej odczytywany w programie MATLAB, czy też po pewnych kolejnych tekstowych przekształceniach także w programach: **WEKA i platformie - R** celem dokonania analizy statystycznej

6. Na poziomie programu **MATLAB -7** autorzy opracowali własny zestaw skryptów z algorytmami do analizy danych pozwalający w szczególności na: poszukiwanie reguł decyzyjnych tkwiących w badanym zbiorze (i liczbową ocenę jakości tych reguł), poszukiwanie tzw. **reguł Pareto - optymalnych**, budowę klasyfikatorów regułowych, podstawową analizę rozkładów prawdopodobieństwa w badanym zbiorze danych oraz prowadzić inne badania informatyczno-analityczne- **stosując metodę Kaplan-Maier**

7. Na poziomie programu WEKA konstruowano i testowano różne klasyfikatory na badanym zbiorze, a w szczególności: naiwny klasyfikator Bayesa, klasyfikatory perceptronowe i neuronowe, drzewa decyzyjne, klasyfikator SVM (*Support Vector Machine*). Testowanie wszystkich klasyfikatorów odbywało się wg mechanizmu tzw. ***krzyżowej walidacji.***

TEORIA-OPIS PROCEDURY ANALITYCZNO- INFORMATYCZNEJ-CD

8. Na poziomie programu platformy- **R** stawiano testy statystyczne dla badanego zbioru, w szczególności: budowano różne tabele kontygencji, obliczano tzw. iloraz szans (*odds ratio*) stosowany powszechnie w medycynie i stawiano związany z nim test Fishera dla małych liczb lub test $\chi 2$ dla dużych liczb badanych chorych.

9. Należy zaznaczyć, że w trakcie badań różne ze zmiennych traktowane były kolejno jako zmienna decyzyjna (inaczej: objaśniana bądź prognozowana), m.in.: status pacjenta (zdrowy / chory), rozpoznanie i typ nowotworu, wiek zachorowania, przedziały wiekowe, okresy przeżycia do wystąpienia zachorowania lub występowanie pewnej mutacji oraz czy istnieje jakiś związek mutacji z pojawieniem się określonego typu zachorowania na raka.

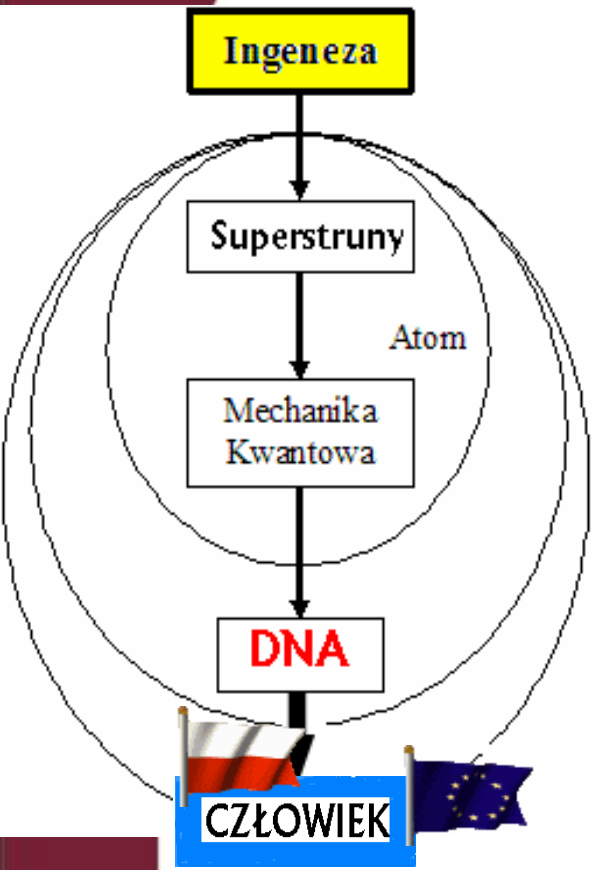
10. W ten sposób sprawdzano, które zmienne dają się najlepiej przewidywać na podstawie pozostałych dostępnych zmiennych lub jak należy dobierać parametry aby uzyskany wynik najlepiej określał stan zdrowia pacjenta. Ale w niniejszej pracy uwzględniono tylko po pierwsze wyłącznie analizy statystyczne dla oceny ryzyka zachorowania na raka u kolejnych pacjentek z mutacją CHEK2 w określonych przedziałach wiekowych, które już zachorowały na raka sutka i porównano to z grupa kontrolną pacjentek z mutacją CHEK2 u których nie stwierdzono nowotworu w tych samych przedziałach wiekowych oraz po drugie stosując różne inne analizy porównywano ich przydatność w naszych analizach dla oceny wystąpienia związku pomiędzy różnymi mutacjami CHEK2 w danej rodzinie a wystąpieniem zachorowania na raka sutka lub wystąpieniem innych nowotworów u członków rodzin będących krewnymi pierwszego rzędu w analizowanej rodzinie

11. W dalszej części niniejszej pracy omówione zostaną zagadnienia medyczne związane z problemem penetracji i genotypowania nowotworowego, a następnie omówione zostaną uzyskane wyniki z badań przeprowadzonych wg wyżej opisanej procedury jak również dla porównania także innymi metodami

POTRZEBY ogólne

Ingeneza, Superstruny, Atom, Kwant, Mechanika Kwantowa, DNA, Mózg-Umysł.

CZŁOWIEK CHCE SŁUSZNIE WIEDZIEĆ CORAZ WIĘCEJ A WSZYSTKO CO DOSKONAŁE DOJRZEWA POWOLI DLATEGO:



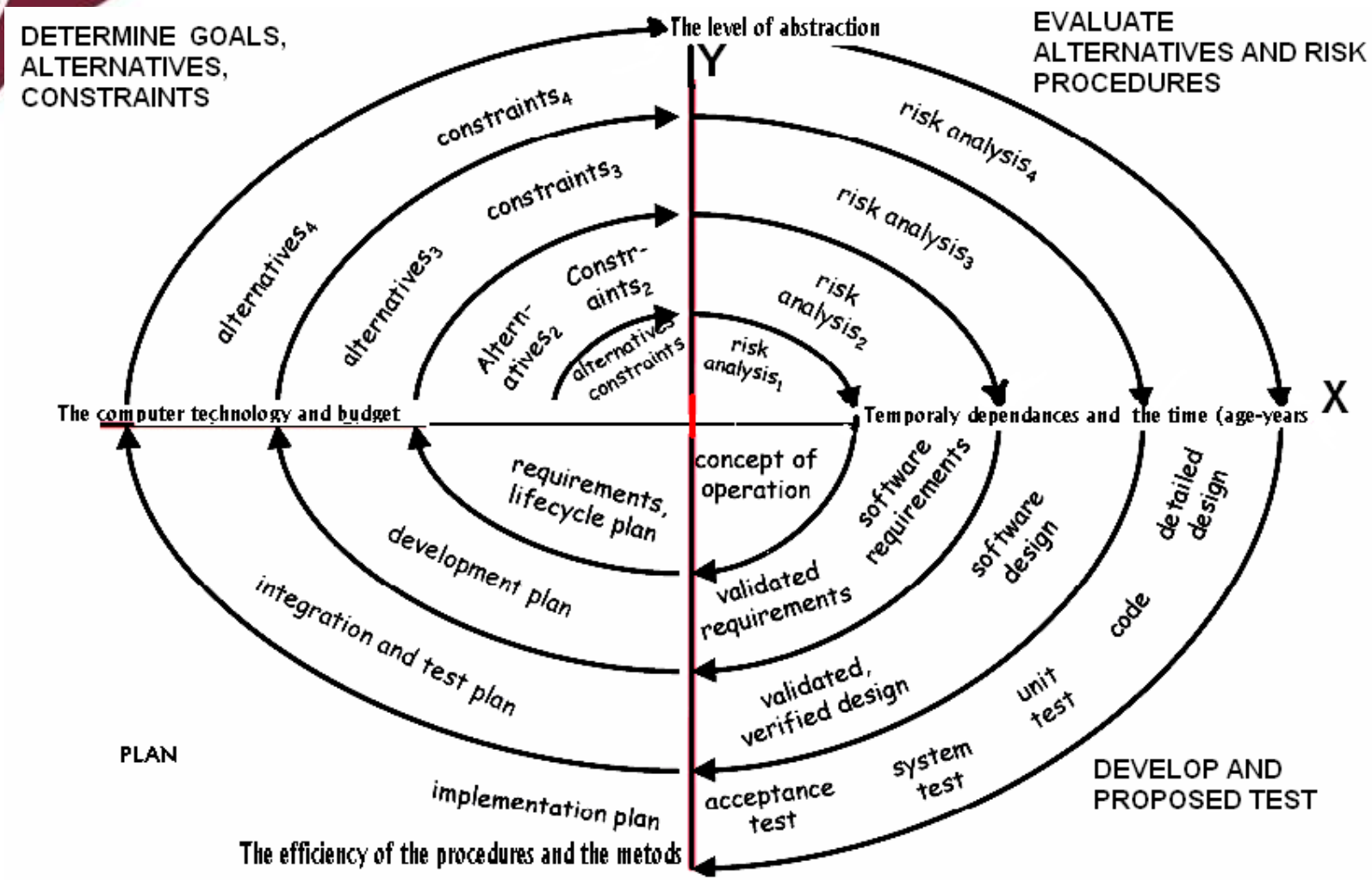
Rys.5

- Jeśli chcemy **rozwiązać** problem naukowy / odnosić sukcesy
- są nam **niezbędne**: a) ciężka praca i światły umysł, b) mieć szczęście do Szefa oraz Współpracowników, c) właściwe metodologie i technologie, d) nowoczesne zaplecze aparaturowe: pyrosequencing, sequenom, bioroboty i oprogramowania, nowoczesne systemy analityczno-informatyczne oraz umiejętności programowania, **e) bioinformatyka (genomika, proteomiką, metabolomika, transkryptomika),** f) szeroko rozumiana współpraca

POTRZEBY dla postępu w GENETYCE

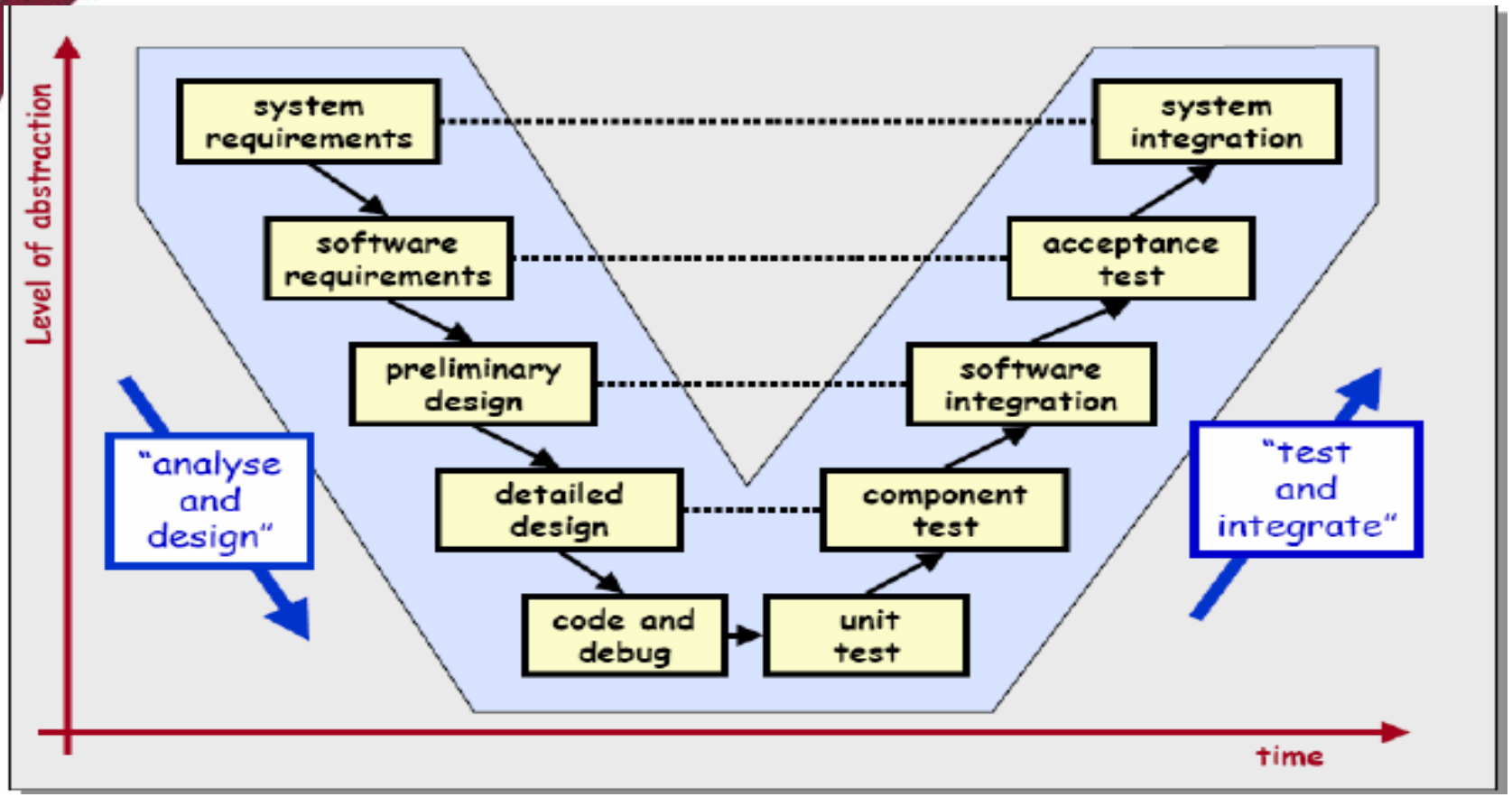
1. **Szczegółowe dane rodowodowe-kliniczne (**rodzinne rodowody) - pełna wiedza o rodzinach oraz o czynnikach sprzyjających rozwojowi choroby (genetyka kliniczna, molekularna, populacji, środowiska oraz poradnictwo genetyczne a także psychogenetyka
2. **GENOMIKA**- funkcjonalna, strukturalna, teoretyczna, porównawcza, indywidualnych różnic (DNA - sekwencjonowanie i komputerowe nanotechnologie oraz biotechnologiczne nanomateriały). Głównym celem genomiki jest poznanie sekwencji materiału genetycznego oraz mapowanie genomu ale również określenie wszelkich zależności i interakcji wewnątrz genomu .
3. **Zrozumienie mechanizmów chorobowych** u człowieka oraz wiedza o mechanizmach regulacji w organizmie i zmienności genetycznej środowiska i na ziemi oraz meta i mikrogenomika mikrobiologiczna.
4. **Czynniki modyfikujące genów** (czynniki wpływające na **ekspresję genów**): metylacja, fosforylacja, acetylacja, zachodzące w aminokwasach końcowych histonów itp.
5. **Ryzyko wystąpienia choroby (OR -Odds ratio**) i **PENETRACJA** nowotworów oraz możliwości zapobiegania chorobie
6. Możliwości finansowe kraju oraz pacjenta
7. Poziom i jakość kadry naukowej - zaangażowanie ludzi i przygotowanie technologiczno-zawodowe
8. Gromadzenie i praktyczne wykorzystywanie informacji naukowych
9. **Promocja Zdrowia Publicznego** oraz postępowanie lekarskie prozdrowotne i nakierowane na sukces (EBM- ***Evidence Based Medicine)***.
10. **Nowoczesna wiedza** o raku, oraz **regulacji epigenetycznej** (procesy regulacji genów niezależne od sekwencji DNA, metylacja sekwencji promotorowych i regulatorowych genów, ubikwitynacja białek). Jednym z kluczowych enzymów jest HADC-deacetylaza histonu

PRZYSZŁOŚĆ -Zaawansowane Systemy Technologiczno - Informatyczne



Rys. 6 Spiralny model proponowanych czynności dla przygotowania procedur i platformy analizy badawczej w skali zaawansowanej - wysoko zaawansowany silnik informatyczno - analityczny stanowiący poważne zaplecze dla planowania bardzo zaawansowanych analiz statystycznych.

MODEL GRAFICZNY -Inżynieria Budowy Oprogramowania



Rys. 7 requierments ↔ design ↔ code ↔ test ↔ integrete ↔ maintain
wymagania ↔ projekt ↔kod↔ testy↔integracja ↔utrzymanie

ESENCJA MODELOWANIA BIOINFORMATYCZNEGO

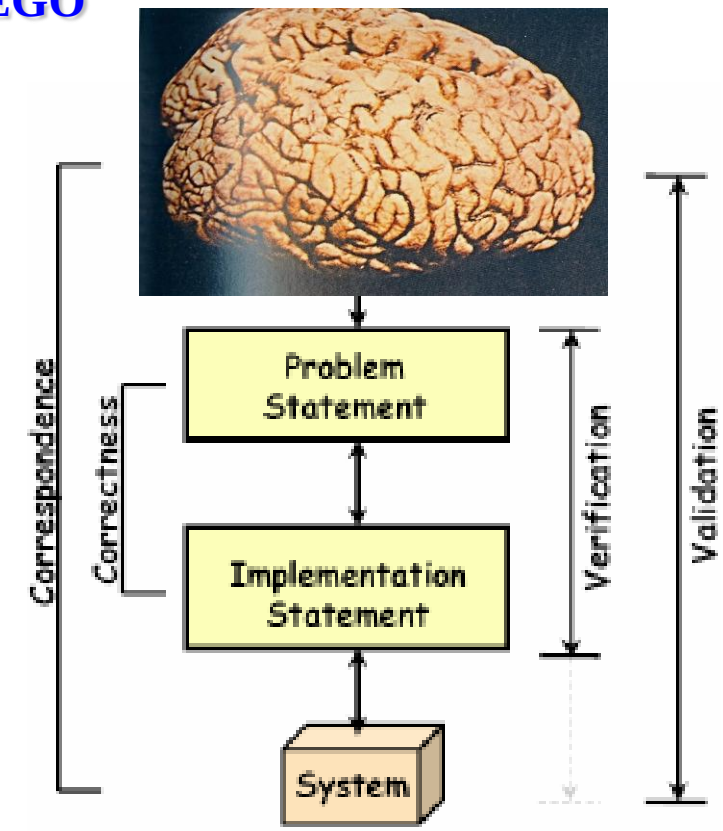
1. Świat rzeczywisty (dotyczy przyszłości człowieka)- Real World
2. Postawienie problemu (wymaga zaawansowanych technologii i umiejętności)- Problem Statement and Contents Statement
3. Implementacje rozwiązania (wymaga zastosowania do realizacji zaawansowanych systemów informatyczno- analitycznych)- Implementation Statement
4. Nowoczesne Zintegrowane Systemy Informatyczne –(Modern Computer - Intergrated Informatics Analyse Statement).
5. Poprawność systemu- Testowanie wszystkich klasyfikatorów wg mechanizmu tzw. *krzyżowej walidacji gdzie podstawową oceny jest zaufanie do reguły równe*

   $P(Y = y \mid X = x)$ - Correctness Statement
6. Odpowiedzialność za wykonywane analizy- (Responsibility Statement)
7. Odbiór i opis ilościowy- weryfikacja problemu- (Verifications Problem)
8. Odbiór i opis jakościowy -walidacja –(Validation Problem)
9. Promocja Zdrowia Publicznego, nowoczesnej wiedzy oraz odpowiedzialnego systemu wartości (The usefulness of computer technologies in the research of the Public Health and in the responsible value-system)
10. **Postępy w bioinformatyce i rozwój nanotechmologii (multicomputer system and nanotechnology and nanoelectronics)**

MODEL GRAFICZNY - ESENCJA MODELOWANIA BIOINFORMATYCZNEGO

- Świat rzeczywisty
- Postawienie problemu
- Implementacja rozwiązania
- System informatyczny

- Poprawność
- Odpowiedniość
- Weryfikacja
  - odbiór ilościowy
- Walidacja
  - odbiór jakościowy



Rys. 8- Graficzne przedstawienie esencji modelowania informatycznego po uwzględnieniu inżynierii oprogramowania w małej skali jaka została zastosowana w naszej propozycji:

requierments ↔ design ↔ code ↔ test ↔ integrete ↔ maintain

(wymagania ↔ projekt ↔kod ↔testy↔integracja ↔utrzymanie)

REGUŁY PARETO - OPTYMALNE

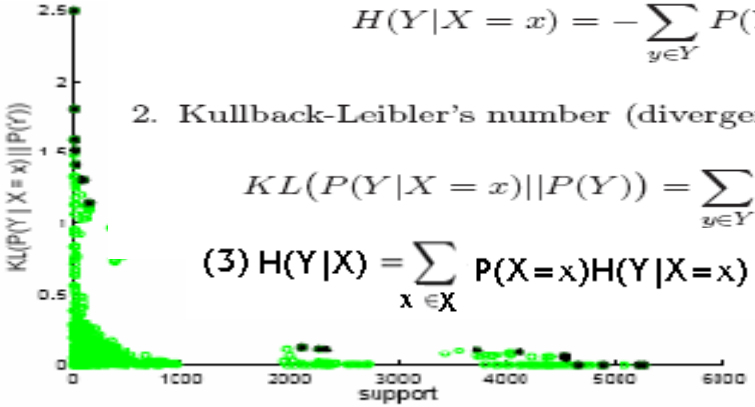
1. Conditional entropy given the fixed $X = x$:

$$H(Y|X = x) = -\sum_{y \in Y} P(Y = y|X = x) \log_2 P(Y = y|X = x).$$

2. Kullback-Leibler's number (divergence):

$$KL(P(Y|X = x)||P(Y)) = \sum_{y \in Y} P(Y = y|X = x) \log_2 \frac{P(Y = y|X = x)}{P(Y = y)}.$$

(3) $H(Y|X) = \sum_{x \in X} P(X=x) H(Y|X=x)$ (4) $KL(Y|X||Y) = \sum_{x \in X} P(X=x)\, KL(Y|X=x||Y)$



Graph. 9 Figure of the Pareto-optimal rules .Rules at Pareto border marked as asterisks

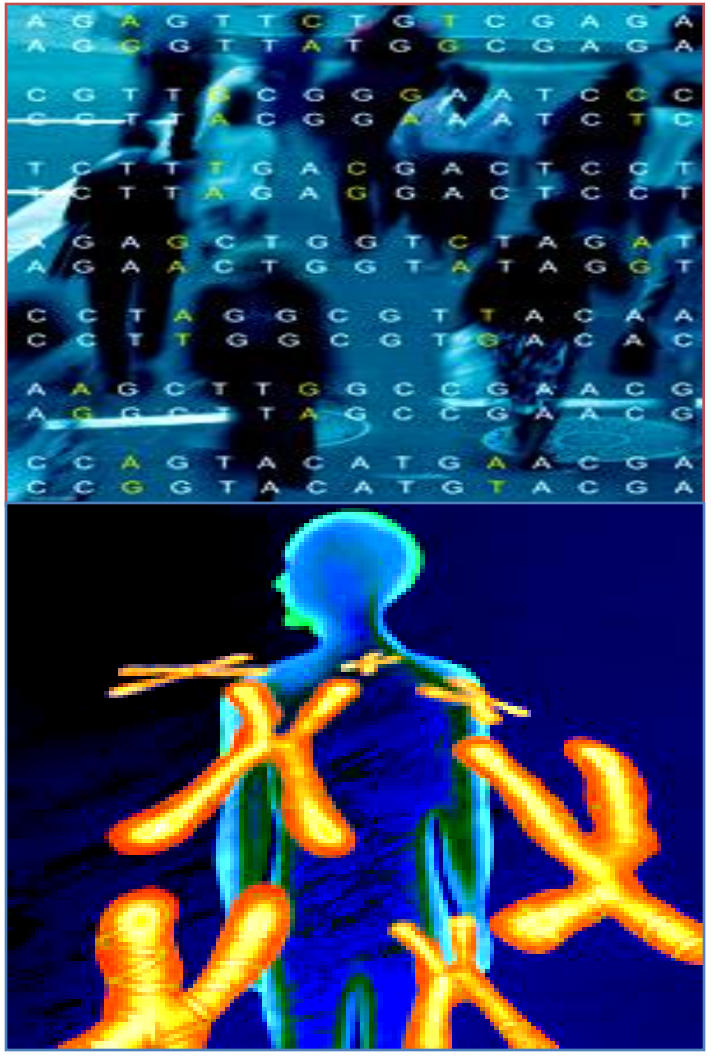
X-oznacza nazwy zmiennych w przesłance. x- wartości przypisane do tych zmiennych w przesłance
Y- nazwa zmiennej decyzyjnej. y- wartość przypisane do tej zmiennej. W przypadku entropii warunkowej, im mniejsza wartość entropii tym lepsza reguła. W przypadku liczby Kullbacka-Leiblera(dewergencji= pseudoodległość rozkładów) im większa ta liczba tym lepsza jest reguła .

Rys.9. Indukcja reguł decyzyjnych dla klasyfikacji-**opracowanie własne.**

Reguły Pareto - optymalne wg tzw. ***krzyżowej walidacji gdzie podstawową oceny jest zaufanie do reguły równe P (Y= y| X= x)*** - : badania na przykładzie medycznego zbioru danych dotyczącego mutacji DNA krwi i nowotworów. Każdej regule należy przyporządkować pewną wartość liczbową, która ocenia, na ile dana reguła jest interesująca oraz jak istotna jest ona statystycznie. Miara *zaufania* w parze ze *wsparciem* (liczbą przypadków w całym zbiorze danych zgodnych z przesłanką reguły) stanowią podstawę znanego algorytmu *A priori*, który stosuje się do poszukiwania reguł asocjacyjnych w dużych zbiorach danych

Znaczenie badań bioinformatycznychtych – korzyści w leczeniu

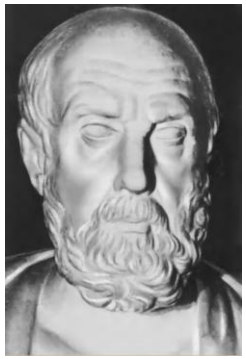
- ograniczenie ryzyka działań niepożądanych
- ograniczenie kosztów leczenia
- zwiększenie skuteczności leczenia
- badanie genetyczne - jeden raz w życiu, czas i
- koszty są usprawiedliwione
- zwłaszcza w przypadku leczenia długotrwałego i skutecznego
- interakcje leków możliwe do przewidzenia



Znaczenie badań bioinformatycznych – korzyści dla medycyny jako nauki – (multicomputer system and nanotechnology)

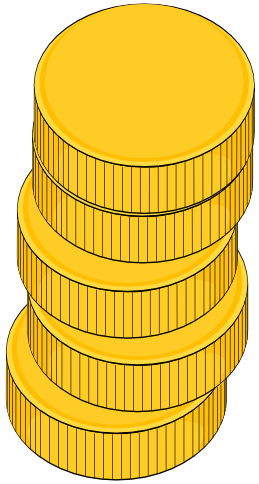
1. katalogowanie informacji biologicznych (bazy danych, *bazy danych sekwencji* i wyszukiwanie sekwencji, anotacji, danych numerycznych w bazach danych)
2. analiza sekwencji DNA (składanie sekwencji, anotacja, wyszukiwanie sekwencji kodujących, regulatorowych i repetytywnych, motywów, markerów)
3. analiza sekwencji genomów, *porównywanie genomów*
4.ustalanie ewolucyjnych relacji pomiędzy zbiorami sekwencji /organizmów (drzewa filogenetyczne) genotypowane (używane między innymi do wyszukiwania genów odpowiedzialnych za *choroby genetyczne*, w ustalaniu ojcostwa, kryminalistyce)
5. analiza ekspresji genów (głównie analiza danych z mikromacierzy)
6. analiza sekwencji *białek*, nazywana też proteomiką (porównywanie sekwencji, wyszukiwanie domen i motywów, przewidywanie własności fizyko-chemicznych, drugo - i trzecio – rzędowej struktury białka, lokalizacji w obrębie komórki, analiza danych z eksperymentów spektroskopowych)
7. katalogowanie funkcji genów/białek, analiza dróg metabolicznych (Np metabolizm lipidów) oraz dróg sygnałowych (Np. od receptora na powierzchni komórki poprzez kaskadę kinaz do czynników transkrypcyjnych)
8. modelowanie układów biologicznych (Np. kinetyka szeregu reakcji enzymatycznych w komórce)
9. wirtualne dokowanie (ang. virtual docking) - Np. używając trójwymiarowej struktury aktywnego centrum enzymu ("zamek" albo "kieszonka" ang. pocket) przeszukuje się w komputerze tysiące małych cząsteczek z których kilka-kilkanaście ('kluczy') będzie miało kształt mieszczący się w centrum aktywnym. Pierwszy krok w kierunku odkrywania nowych leków.
10. **BIOKOMPUTERY *DNA*-** w którym obliczenia zachodzą dzięki reakcjom chemicznym między cząsteczkami *DNA* oraz możliwa jest do wykonania **morfometria** / *analiza obrazu*

Uzasadnienie dla BIONFORMATYKI



Zasada Hipokratesa – *Primum non nocere -* *lekarska zasada*

BIONFORMATYKA - to stworzenie możliwości zmniejszania kosztów farmakoterapii oraz zastosowanie innych nowoczesnych, naukowo uzasadnionych procedur medycznych według zasady ALARA (As Low As Reasonably Achievable) pozwalające na racjonalizację wydatków przeznaczonych na opiekę zdrowotną, skuteczne prowadzenie badań naukowych oraz promocję ZDROWIA PUBLICZNEGO – propozycja własna



ALARA – tak mało, jak to jest rozsądnie osiągalne

WNIOSKI

1. Własny autorski program-pomysł do analiz statystycznych oraz analizy ryzyka zachorowania na raka sutka jest porównywalny z innymi programami oraz może być przydatny i z powodzeniem stosowany do wykonywania zaawansowanych analiz statystycznych oraz liczenia penetracji różnych typów raka i mutacji w zakresie krewnych pierwszego rzędu jak również w szerszym zakresie pod warunkiem prawidłowego wykonania rodowodu i zastosowania wszystkich elementów opisanej procedury.
2. Zaproponowany sposób i metoda pozwala na prawie automatyczne stworzenie bazy danych probantów lub nosicieli mutacji w oparciu o struktury rodowodowe oraz następnie przygotowane dane mogą być poddane każdej analizie statystycznej w oparciu o którą można wykonywać specjalistyczne analizy statystyczne a także liczyć penetrację dla określonej mutacji lub zachorowania
3. Metoda-pomysł jest bardzo efektywna, tania, eliminuje **dublowanie** czynności, jest szybka i porównywalna z wysokospecjalistycznymi systemami typu SAS lub statystyki proponowane przez StatSoft Polska typu STATYSTYKA-8.0 lub STATYSTYKA –Data Miner oraz posiada wysoką specyficzność i dokładność oraz pozwala na szybkie przygotowanie bazy danych do analiz statystyczno-obliczeniowych oraz liczenia penetracji lub wykonywania innych obliczeń dla oceny ryzyka badanej choroby lub procesu oceny zdrowia badanej populacji.
4. Autor nie podważa skuteczności stosowania innych procedur analitycznych w obliczeniach statystycznych, jednak jest przekonany o wysokiej skuteczności proponowanego rozwiązania, pozwalającego na bardzo szybkie i tanie uzyskanie przydatnej i praktycznej wiedzy dla zaleceń lekarskich.
5. Wykorzystując przez wiele lat gromadzony w Zakładzie Genetyki PAM i posiadany materiał w postaci rodowodów rodzinnych można bardzo szybko przygotować do analizy po przekształceniu rodowodów z postaci graficznej do postaci tekstowej - tabeli i po uwzględnieniu opisanych procedur badawczych oraz w zależności od ilości probantów i badanych cech wykonywać bardzo precyzyjne i bardzo zaawansowane analizy statystyczne.

Welcome for collaboration.

Zakład Genetyki i Patomorfologii PAM w Szczecinie, ul. Połabska 4, 71-115 Szczecin, Kierownik. Profesor dr hab. med. Jan Lubiński

More information:

☞ www.hereditarycancer.netemail:ihcc@wp.pl
☞ phone:+48-91-466-15-42

fax: +48-91-466-15-33

DZIĘKUJĘ PAŃSTWU ZA UWAGĘ

Email :(cf.segment@sci.pam.szczecin.pl)