

Antal Péter – Arany Ádám – Bolgár Bence – Gézsi András – Hajós Gergely
 – Hullám Gábor – Marx Péter – Millinghoffer András – Poppe László
 – Sárközy Péter

BIOINFORMATIKA: MOLEKULÁRIS MÉRÉSTECHNIKÁTÓL AZ ORVOSI DÖNTÉSTÁMOGATÁSIG

A molekuláris biológiai méréstechnikai fejlődés a nagy adattömegeket, majd a hipotézismentes kutatási paradigma megjelenését hozta el az orvosbiológiába. Az ezredforduló előtti genetikai-genomikai korszakot a posztgenomikai korszak követte egyre szaporodó omikai szintekkel és leíró hálózati megközelítésekkel. Egy évtized után azonban egyre inkább a nagyléptékű adat- és tudásfúzió került a központba. A jegyzet ezen új kihívásokat tekinti át. Az első két fejezet a genetikai méréstechnika alapjait foglalja össze. A genetikai variánsok hatásainak megértését a fehérjék szerkezetének tárgyalása, ill. a génszabályozási hálózatok bemutatása segíti a következő két-két fejezetben. Ezután az alapvető fontosságú statisztikai asszociációs elemzéseket mutatja be. Az értelmezés támogatására összefoglaljuk az oksági következtetés egy Bayes-hálókon alapuló formalizálását, ill. a szövegbányászati módszereket. A kísérletek szekvencialitása mellett az adatok heterogenitása és így integrált elemzése is központi kihívás, amely kihívást még nehezebbé tesz az egyre elérhetőbb „mély”, azaz részleteiben gazdag fenotípus- és környezeti leírások. Az adatmegosztás hatékonysága miatt és a nagy számításigény miatt is egyre fontosabbá válnak az általánosan elérhető, közmű jellegű informatikai szolgáltatások, amelyek működését példákkal is illusztráljuk. Az áttekintést egy gyógyszerkutatási összefoglaló zárja, amelyben a személyre szabott medicina szempontjai is megjelennek, ill. egy metagenomikai összefoglaló, amely az epigenetikai szint megjelenése után korunk egy új ígéretes omikai szintje.

Kulcsszavak: genotipizálás, új generációs szekvenálási módszerek, fehérjemodellezés, génszabályozási hálózatok, omikai hálózatok, dinamikus rendszerek, kísérlettervezés, munkafolyamatrendszerek, asszociációs elemzések, biomarker-elemzések, adat- és tudásfúzió, oksági következtetés, orvosi döntéstámogató rendszerek, nagy adattömegek, szemantikus publikálás, hasonlósági alapú gyógyszerkutatás, metagenomika.

Budapesti Műszaki és Gazdaságtudományi Egyetem és Semmelweis Egyetem



Typotex Kiadó
 2014

COPYRIGHT: © 2014–2019, Antal Péter, Arany Ádám, Bolgár Bence, Gézsi András, Hajós Gergely, Hullám Gábor, Marx Péter, Millinghoffer András, Poppe László, Sárközy Péter, Budapesti Műszaki és Gazdaságtudományi Egyetem, Semmelweis Egyetem

Creative Commons NonCommercial-NoDerivs 3.0 (CC BY-NC-ND 3.0)

A szerző nevének feltüntetése mellett nem kereskedelmi céllal szabadon másolható, terjeszthető, megjelentethető és előadható, de nem módosítható.

Szakmai lektorok: Molnár Viktor, Antos András

ISBN 978 963 279 180 7

Készült a Typotex Kiadó gondozásában

Felelős vezető: Votisky Zsuzsa

Készült a TÁMOP-4.1.2/A/1-11/1-2011-0079 számú, „Konzorcium a biotechnológia aktív tanulásáért” című projekt keretében.

Nemzeti Fejlesztési Ügynökség
www.ujszachenyiterv.gov.hu
06 40 638 638



A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Tartalomjegyzék

1. DNS rekombináns méréstechnológiák, zaj- és hibamodellek	11
1.1. Történelmi áttekintés	11
1.1.1. A genomszekvenálás klinikai aspektusai	12
1.1.2. Részleges genetikai asszociációs vizsgálatok (PGAS)	12
1.1.3. Genomszintű asszociációs vizsgálatok (GWAS)	12
1.2. Első generációs automatizált Sanger-szekvenálás	13
1.3. Új generációs szekvenálási technológiák	13
1.3.1. Piroszekvenálás és pH alapú szekvenálás	13
1.3.2. Reverzibilis terminátor alapú szekvenálás	15
1.3.3. Nanopórus alapú szekvenálás	16
1.4. Új generációs szekvenálási technológiák hibakarakterisztikája	17
1.4.1. Carry forward/incomplete extension	18
1.4.2. Homopolimer hibák	18
1.5. Capture technológiák	19
1.5.1. PCR capture	19
1.6. Emulziós PCR	22
1.7. Híd- (bridge-) amplifikáció	23
1.8. Célzott újraszekvenálás	23
1.9. De novo szekvenálás	24
1.10. Új generációs szekvenálási munkafolyamatok	24
1.10.1. Szűrés	24
1.10.2. Illesztés	24
1.10.3. Összerakás	24
1.10.4. Variánshívás	25
1.10.5. Paired-end szekvenálás	25
1.11. Több minta párhuzamos szekvenálása	26
2. Genetikai mérések és utófeldolgozásuk, haplotípus-rekonstrukció, impu- tálás	27
2.1. A genom fogalma	27
2.2. A genotípus „az egyed genetikai identitása”	28
2.2.1. Egy pontos nukleotid-polimorfizmus (SNP)	29
2.2.2. A pontmutációk lehetséges változatai	29

2.2.3.	Mutációk hatása	30
2.3.	Haplotípusok	31
2.4.	Kapcsoltsági egyensúlytalanság	31
2.5.	Haplotípus-rekonstrukció	32
2.6.	Imputálás	34
2.7.	Genotipizálási módszerek	35
2.7.1.	Sanger-szekvenálás	36
2.7.2.	Valós idejű kvantitatív PCR	36
2.7.3.	DNS chipok	36
2.8.	Genotipizálás és génexpresszió	38
2.8.1.	Sikeres mérések és pontosságuk	38
3.	Összehasonlító fehérjemodellezés és molekuladokkolás	39
3.1.	Bevezetés	39
3.1.1.	A fehérjeszekvencia-szerkezeti szakadék	40
3.1.2.	A fehérjemodellezés módszerei	41
3.2.	Összehasonlító fehérjemodellezés	42
3.2.1.	A homológiamodellezés lépései	42
3.2.2.	Homológiamodellezési eszközök	47
3.3.	Molekuladokkolás	49
3.3.1.	Fehérje–ligandum kölcsönhatás-előrejelzések	50
3.3.2.	Fehérje–biomakromolekula kölcsönhatás-előrejelzések	51
4.	Fehérjeszerkezet-meghatározás kísérleti módszerei és egyszerű fehérje-szerkezet-predikciók	56
4.1.	Bevezetés	56
4.1.1.	A fehérjeazonosítás eszközei	56
4.1.2.	Egyszerű fehérjeanalízis	57
4.1.3.	A fehérjeszerkezet-előrejelzés szintjei és nehézségei	57
4.2.	Fehérjék másodlagos szerkezetének kísérletes vizsgálata	58
4.2.1.	Fehérje cirkuláris dikroizmus (CD)	59
4.2.2.	Szinkrotron besugárzásos cirkuláris dikroizmus (SRCD)	60
4.2.3.	Kísérleti módszerek fehérjék atomi szintű szerkezetének meghatározására	60
4.2.4.	Fehérje-röntgenkristallográfia	62
4.2.5.	Fehérje-NMR-spektroszkópia	63
4.2.6.	Fehérje-elektronmikroszkópia, elektrondiffrakció és elektronkristallográfia	66
4.2.7.	Fehérje-neutronkristallográfia	67
5.	Genetikai variánsok funkcionális hatásainak kvantitatív modelljei	70
5.1.	Bevezetés	70
5.2.	Variánsok	70

5.2.1.	SNP, indel	71
5.2.2.	Alternatív splicing	72
5.3.	A szabályozás szintjei	72
5.4.	Különböző szabályozó elemek	72
5.5.	microRNS	72
5.5.1.	miRNS érés	73
5.5.2.	miRNS által mediált szabályozási formák	73
5.6.	Transzkripciós faktorok	74
5.7.	Epigenetika	74
5.7.1.	Metiláció	75
5.7.2.	Hisztónmódosulások	75
5.8.	Modellezés	76
5.8.1.	regSNP	76
5.8.2.	Boolean modellek	76
5.8.3.	Termodinamikai modellek	77
5.8.4.	Differenciálegyenletek	77
5.8.5.	Lac operon	78
6.	Génszabályozási hálózatok matematikai modelljei	82
6.1.	Bevezetés	82
6.2.	Hálók tanulása	82
6.3.	Nem felügyelt tanulási módszerek	83
6.3.1.	ARACNE	84
6.3.2.	REVEAL	84
6.4.	Felügyelt módszerek	85
6.4.1.	PosOnly	86
6.4.2.	SIRENE	86
6.5.	TF, miRNS, mRNS szabályozó hálózatok	87
7.	Genetikai asszociációs vizsgálatok standard elemzése	90
7.1.	Bevezetés	90
7.2.	Genetikai adattranszformáció	91
7.2.1.	Szűrés	91
7.2.2.	Hardy–Weinberg-egyenlőség vizsgálata	91
7.3.	Fenotípus-adattranszformáció	92
7.3.1.	Transzformáció	93
7.3.2.	Diszkretizálás	93
7.4.	Egyváltozós statisztikai módszerek	93
7.4.1.	Standard asszociációs tesztek	93
7.4.2.	Cochran–Armitage-trendteszt	96
7.4.3.	Hatáserősség	97
7.4.4.	Egyváltozós Bayes-i módszerek	98
7.5.	Többváltozós módszerek	99

7.5.1.	Logisztikus regresszió	99
7.5.2.	Haplotípus-asszociáció	100
7.5.3.	Statisztikai erő vizsgálata	104
8.	Génexpressziós adatok standard asszociációs elemzése	107
8.1.	Bevezetés	107
8.2.	Előfeldolgozás	108
8.2.1.	Háttérkorrekció	108
8.2.2.	Normalizáció	109
8.2.3.	Összegzés	109
8.2.4.	Szűrés	110
8.3.	Adatelemzés	111
8.3.1.	Klaszterezés	111
8.3.2.	Differenciális expresszió	115
8.3.3.	Az eredmények biológiai értelmezése	116
9.	Biomarker-elemzés	121
	Jelölések	121
9.1.	Bevezető	123
9.2.	Elméleti háttér	124
9.3.	Bayes-i többszintű relevancia-elemzés	127
9.4.	Többváltozós skálázhatóság: a k-MBS jegy	128
9.5.	Többcélváltozós relevancia	130
9.6.	Poszterior-dekomponáláson alapuló interakció és redundancia	130
9.7.	MBS poszteriorok utófeldolgozása és megjelenítése	131
9.8.	Tudás alapú utóaggregálás	132
9.9.	Összefoglaló	132
10.	Hálózatbiológia	135
10.1.	Bevezetés	135
10.2.	Biológiai hálózatok	136
10.3.	Gráfelméleti alapok	137
10.4.	Hálózatelemzés	138
10.4.1.	Hálózati topológia	138
10.4.2.	Hálózati modellek és dinamika	139
10.4.3.	Asszortativitás, fokszámeloszlás és skálafüggetlen hálózatok	140
10.4.4.	Feladatok és kihívások	141
10.5.	Néhány alkalmazás	143
11.	Dinamikus modellezés a sejtbiológiában	147
11.1.	Biokémiai fogalmak, ezek számításhoz való reprezentációi	147
11.2.	Modellezés differenciálegyenletekkel	150
11.3.	Sztochasztikus modellezés	151

11.4. Hibrid módszerek	152
11.5. Reakció–diffúzió-rendszerek	153
11.6. Modell-illesztés	154
11.7. Teljes-sejt-szimuláció	155
11.8. Áttekintés	156
12. Oksági következtetések az orvosbiológiában	158
Jelölések	158
12.1. Bevezető	160
12.2. Függetlenségi és oksági relációk reprezentálása Bayes-hálókkal	161
12.3. Oksági relációk kényszer alapú tanulása	165
12.4. Teljes oksági modellek Bayes-i tanulása	166
12.5. Oksági jegyek következtetése Bayes-halók feletti átlagolással	167
12.5.1. Élek: közvetlen páronkénti függések	168
12.5.2. Áttételes páronkénti oksági relációk	169
12.5.3. Markov-takaró (al)gráf	169
12.5.4. Hatásmódosítók	170
12.5.5. Változók sorrendje	171
13. Szövegbányászati módszerek a bioinformatikában	174
13.1. Bevezetés	174
13.2. Orvosbiológiai szövegbányászat	174
13.2.1. Korpuszépítés	175
13.2.2. Szótárépítés	177
13.2.3. Szövegbányászati feladatok	178
13.3. Alapvető szövegbányászati technikák	179
13.3.1. Mintaillesztés	179
13.3.2. Dokumentumok reprezentációja	179
13.3.3. Az entitásfelismerés módszerei	181
13.3.4. A relációkivonatolás módszerei	182
13.3.5. Lexikalizált valószínűségi környezetfüggetlen nyelvtanok	183
13.3.6. Az orvosbiológiai szövegbányászat kihívásai	184
13.4. Szövegbányászat és tudásszerzés	185
14. Kísérlettervezés: az alapoktól a tudásgazdag és aktív tanulós kiterjesztésekig	188
14.1. Bevezetés	188
14.2. A kísérlettervezés alapjai	188
14.2.1. Az orvosbiológiai kísérlettervezés lépései	189
14.2.2. A biológiai kísérletek fajtái	189
14.3. A kísérlettervezés döntéelméleti megközelítése	191
14.3.1. A kísérlet várható értéke	191
14.3.2. Adaptív kísérlettervezés és költségkorlátozott tanulás	193

14.3.3. Szekvenciális döntési folyamatok Bayes-i keretben	194
14.4. A célváltozók kiválasztását szolgáló módszerek	195
14.4.1. Géprioritizálás	195
14.4.2. Aktív tanulás	197
14.5. Egyéb, a gyakorlatban felmerülő bioinformatikai feladatok	198
15. Nagy adattömegek az orvosbiológiában	201
15.1. Bevezető	201
15.2. Az orvosbiológia klasszikus nagy adattömegei	202
15.3. Posztgenomikai nagy adattömegek az orvosbiológiában	203
15.4. Hétköznapiakból származó nagy adattömegek	206
15.5. A hétköznapi nagy adattömegek az orvosbiológiában	208
15.6. A hétköznapi nagy adattömegek bioinformatikai kihívásai	211
16. Heterogén biológiai adatok fúziós elemzése	216
16.1. Bevezetés	216
16.2. Tudásfúzió és adatfúzió	218
16.3. Az adatfúzió módszereinek felosztása	219
16.3.1. Korai fúzió	220
16.3.2. Köztes fúzió	221
16.3.3. Késői fúzió	221
16.4. Hasonlóság alapú adatfúzió	222
17. A Bayes-i enciklopédia	227
17.1. Bevezető	227
17.2. Az adat, tudás, számítás hármának modern kori megjelenései	231
17.3. Az adat, tudás, számítás hármása a genetikai asszociációs kutatásokban	232
17.4. Trendek az adatvilágban	234
17.4.1. Új generációs szekvenálási adatok feldolgozásának dokumentálása	235
17.4.2. Gazdag fenotípusos adatok	235
17.5. Trendek a tudásvilágban: szemantikus publikálás és adatelemzési tudásbázisok	236
17.5.1. Szemantikus publikálás	236
17.5.2. Adatelemzési tudásbázisok	237
17.6. Trendek a modellvilágban	238
18. Bioinformatikai munkafolyamat-rendszerek	
— esettanulmány	243
18.1. A feladat áttekintése	243
18.2. Adatmodell és -reprezentáció	244
18.3. Felhasználói esetek és architektúra	245
18.4. A szerver működési részletei	247
18.5. Utófeldolgozási lépések	248

19.A gyógyszeripari kutatás informatikai aspektusai	250
19.1. A fejlesztési folyamat áttekintése	250
19.2. Kemoinformatikai háttér	251
19.3. Szűrési kritériumok	253
19.4. Módszerek	256
19.5. Fragmens alapú tervezés	259
19.6. Gyógyszer-újrapozicionálás	260
20. Metagenomika	264
20.1. Bevezetés	264
20.2. A metagenom elemzése	265
20.2.1. A közösséget alkotó fajok beazonosítása	265
20.2.2. Funkcionális metagenomika	266
20.3. Metagenomika lépésről lépésre	267
20.3.1. Mintavételezés	267
20.3.2. Szekvenálás	269
20.3.3. Genomösszerakás	269
20.3.4. Besorolás	270
20.3.5. Génfelismerés és funkcionális annotáció	271