

Schedule for coming weeks

- Homework
 - HW 1 graded
 - HW 2 due Wed 11/9
 - HW 3 out Wed 11/16, due Monday 12/5
 - Incremental
- Project
 - First two meetings done
 - Meetings Thursday 11/17
 - Final report/demo due 12/9
- Final 12/8 12-3

Multiple sequence alignment

Bioinformatics Fall 2011

1

Multiple alignment

- An essential tool in molecular biology
 - Finding highly conserved subregions or embedded patterns in a set of biological sequences
 - Conserved regions usually are key functional regions, prime targets for drug developments
 - Estimation of evolutionary distance between sequences
 - Prediction of protein secondary/tertiary structure

Multiple sequence alignment

Bioinformatics Fall 2011

2

```

-----VHLT PEEKGAVTALNGK VM VD --EVGGEALGRLLVY YP VPKR
-----VQLS GEEKAAVGLALNGK VM EE --EVGGEALGRLLVY YP VTQR
-----VLS PADKTNVKAAMGK VG AH AGEYGAEALERDFLS FP TTKT
-----VLS AADKTNVKAAMGK VG GH AGEYGAEALERDFLS FP TTKT
PIVDYGSVAPLS AAEKTYLRSAMAP VY SD YETSVDVILLVFFETS TP AAKE
-----VLS PEGKALVLRHAKM VE AD VAKSRSKDTLRLPKS WP KTLK
-----GALT ESQAALVKSNEE FN AH IPKHTRIFRFLVLEL AP AAKD

FFESFGDLSTFDAYMGN PKVKAHQKVLGAFSDG --L AHEIDL EG TPAI--LSLEIKMLRYD
FFESFGDLSTFDAYMGN PKVKAHQKVLGAFSDG --V AHEIDL EG TPAI--LSLEIKMLRYD
YFPIF-DLSH----GS AQVKAHGKVDALNTA --V ARVDON PG ALSA--LSDLRMLRYD
YFPIF-DLSH----GS AQVKAHGKVDALNTA --V GHELDL PG ALSA--LSDLRMLRYD
FFPKFGLTADLKCS ADVRHHAERATIDAVDA --V ASKMDT EN MSQKLSQKMAKFEVD
KDFRPHLLEADKAS EDLAKHGVTIVLALGAI --L KMKGRH EA ELKP--LAGSARVGRIP
LSSSFLKQTSSEVPMI PELQARAGVFKLYEAA IQE EYTVGV AS DATLQNLGSVEKGVYA
    
```

Alignment between globins (human beta globin, horse beta globin, human alpha globin, horse alpha globin, cyanohaemoglobin, whale myoglobin, leghaemoglobin) produced by Clustal. Boxes mark the seven alpha helices composing each globin.

Multiple sequence alignment Bioinformatics Fall 2011

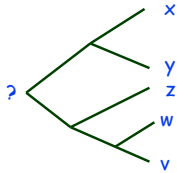
3

Alignment of chromo domains

Classical chromo domain: 13 84 drt... 114... 117... 120... 123... 126... 129... 132... 135... 138... 141... 144... 147... 150... 153... 156... 159... 162... 165... 168... 171... 174... 177... 180... 183... 186... 189... 192... 195... 198... 201... 204... 207... 210... 213... 216... 219... 222... 225... 228... 231... 234... 237... 240... 243... 246... 249... 252... 255... 258... 261... 264... 267... 270... 273... 276... 279... 282... 285... 288... 291... 294... 297... 300... 303... 306... 309... 312... 315... 318... 321... 324... 327... 330... 333... 336... 339... 342... 345... 348... 351... 354... 357... 360... 363... 366... 369... 372... 375... 378... 381... 384... 387... 390... 393... 396... 399... 402... 405... 408... 411... 414... 417... 420... 423... 426... 429... 432... 435... 438... 441... 444... 447... 450... 453... 456... 459... 462... 465... 468... 471... 474... 477... 480... 483... 486... 489... 492... 495... 498... 501... 504... 507... 510... 513... 516... 519... 522... 525... 528... 531... 534... 537... 540... 543... 546... 549... 552... 555... 558... 561... 564... 567... 570... 573... 576... 579... 582... 585... 588... 591... 594... 597... 600... 603... 606... 609... 612... 615... 618... 621... 624... 627... 630... 633... 636... 639... 642... 645... 648... 651... 654... 657... 660... 663... 666... 669... 672... 675... 678... 681... 684... 687... 690... 693... 696... 699... 702... 705... 708... 711... 714... 717... 720... 723... 726... 729... 732... 735... 738... 741... 744... 747... 750... 753... 756... 759... 762... 765... 768... 771... 774... 777... 780... 783... 786... 789... 792... 795... 798... 801... 804... 807... 810... 813... 816... 819... 822... 825... 828... 831... 834... 837... 840... 843... 846... 849... 852... 855... 858... 861... 864... 867... 870... 873... 876... 879... 882... 885... 888... 891... 894... 897... 900... 903... 906... 909... 912... 915... 918... 921... 924... 927... 930... 933... 936... 939... 942... 945... 948... 951... 954... 957... 960... 963... 966... 969... 972... 975... 978... 981... 984... 987... 990... 993... 996... 999... 1002... 1005... 1008... 1011... 1014... 1017... 1020... 1023... 1026... 1029... 1032... 1035... 1038... 1041... 1044... 1047... 1050... 1053... 1056... 1059... 1062... 1065... 1068... 1071... 1074... 1077... 1080... 1083... 1086... 1089... 1092... 1095... 1098... 1101... 1104... 1107... 1110... 1113... 1116... 1119... 1122... 1125... 1128... 1131... 1134... 1137... 1140... 1143... 1146... 1149... 1152... 1155... 1158... 1161... 1164... 1167... 1170... 1173... 1176... 1179... 1182... 1185... 1188... 1191... 1194... 1197... 1200... 1203... 1206... 1209... 1212... 1215... 1218... 1221... 1224... 1227... 1230... 1233... 1236... 1239... 1242... 1245... 1248... 1251... 1254... 1257... 1260... 1263... 1266... 1269... 1272... 1275... 1278... 1281... 1284... 1287... 1290... 1293... 1296... 1299... 1302... 1305... 1308... 1311... 1314... 1317... 1320... 1323... 1326... 1329... 1332... 1335... 1338... 1341... 1344... 1347... 1350... 1353... 1356... 1359... 1362... 1365... 1368... 1371... 1374... 1377... 1380... 1383... 1386... 1389... 1392... 1395... 1398... 1401... 1404... 1407... 1410... 1413... 1416... 1419... 1422... 1425... 1428... 1431... 1434... 1437... 1440... 1443... 1446... 1449... 1452... 1455... 1458... 1461... 1464... 1467... 1470... 1473... 1476... 1479... 1482... 1485... 1488... 1491... 1494... 1497... 1500... 1503... 1506... 1509... 1512... 1515... 1518... 1521... 1524... 1527... 1530... 1533... 1536... 1539... 1542... 1545... 1548... 1551... 1554... 1557... 1560... 1563... 1566... 1569... 1572... 1575... 1578... 1581... 1584... 1587... 1590... 1593... 1596... 1599... 1602... 1605... 1608... 1611... 1614... 1617... 1620... 1623... 1626... 1629... 1632... 1635... 1638... 1641... 1644... 1647... 1650... 1653... 1656... 1659... 1662... 1665... 1668... 1671... 1674... 1677... 1680... 1683... 1686... 1689... 1692... 1695... 1698... 1701... 1704... 1707... 1710... 1713... 1716... 1719... 1722... 1725... 1728... 1731... 1734... 1737... 1740... 1743... 1746... 1749... 1752... 1755... 1758... 1761... 1764... 1767... 1770... 1773... 1776... 1779... 1782... 1785... 1788... 1791... 1794... 1797... 1800... 1803... 1806... 1809... 1812... 1815... 1818... 1821... 1824... 1827... 1830... 1833... 1836... 1839... 1842... 1845... 1848... 1851... 1854... 1857... 1860... 1863... 1866... 1869... 1872... 1875... 1878... 1881... 1884... 1887... 1890... 1893... 1896... 1899... 1902... 1905... 1908... 1911... 1914... 1917... 1920... 1923... 1926... 1929... 1932... 1935... 1938... 1941... 1944... 1947... 1950... 1953... 1956... 1959... 1962... 1965... 1968... 1971... 1974... 1977... 1980... 1983... 1986... 1989... 1992... 1995... 1998... 2001... 2004... 2007... 2010... 2013... 2016... 2019... 2022... 2025... 2028... 2031... 2034... 2037... 2040... 2043... 2046... 2049... 2052... 2055... 2058... 2061... 2064... 2067... 2070... 2073... 2076... 2079... 2082... 2085... 2088... 2091... 2094... 2097... 2100... 2103... 2106... 2109... 2112... 2115... 2118... 2121... 2124... 2127... 2130... 2133... 2136... 2139... 2142... 2145... 2148... 2151... 2154... 2157... 2160... 2163... 2166... 2169... 2172... 2175... 2178... 2181... 2184... 2187... 2190... 2193... 2196... 2199... 2202... 2205... 2208... 2211... 2214... 2217... 2220... 2223... 2226... 2229... 2232... 2235... 2238... 2241... 2244... 2247... 2250... 2253... 2256... 2259... 2262... 2265... 2268... 2271... 2274... 2277... 2280... 2283... 2286... 2289... 2292... 2295... 2298... 2301... 2304... 2307... 2310... 2313... 2316... 2319... 2322... 2325... 2328... 2331... 2334... 2337... 2340... 2343... 2346... 2349... 2352... 2355... 2358... 2361... 2364... 2367... 2370... 2373... 2376... 2379... 2382... 2385... 2388... 2391... 2394... 2397... 2400... 2403... 2406... 2409... 2412... 2415... 2418... 2421... 2424... 2427... 2430... 2433... 2436... 2439... 2442... 2445... 2448... 2451... 2454... 2457... 2460... 2463... 2466... 2469... 2472... 2475... 2478... 2481... 2484... 2487... 2490... 2493... 2496... 2499... 2502... 2505... 2508... 2511... 2514... 2517... 2520... 2523... 2526... 2529... 2532... 2535... 2538... 2541... 2544... 2547... 2550... 2553... 2556... 2559... 2562... 2565... 2568... 2571... 2574... 2577... 2580... 2583... 2586... 2589... 2592... 2595... 2598... 2601... 2604... 2607... 2610... 2613... 2616... 2619... 2622... 2625... 2628... 2631... 2634... 2637... 2640... 2643... 2646... 2649... 2652... 2655... 2658... 2661... 2664... 2667... 2670... 2673... 2676... 2679... 2682... 2685... 2688... 2691... 2694... 2697... 2700... 2703... 2706... 2709... 2712... 2715... 2718... 2721... 2724... 2727... 2730... 2733... 2736... 2739... 2742... 2745... 2748... 2751... 2754... 2757... 2760... 2763... 2766... 2769... 2772... 2775... 2778... 2781... 2784... 2787... 2790... 2793... 2796... 2799... 2802... 2805... 2808... 2811... 2814... 2817... 2820... 2823... 2826... 2829... 2832... 2835... 2838... 2841... 2844... 2847... 2850... 2853... 2856... 2859... 2862... 2865... 2868... 2871... 2874... 2877... 2880... 2883... 2886... 2889... 2892... 2895... 2898... 2901... 2904... 2907... 2910... 2913... 2916... 2919... 2922... 2925... 2928... 2931... 2934... 2937... 2940... 2943... 2946... 2949... 2952... 2955... 2958... 2961... 2964... 2967... 2970... 2973... 2976... 2979... 2982... 2985... 2988... 2991... 2994... 2997... 3000... 3003... 3006... 3009... 3012... 3015... 3018... 3021... 3024... 3027... 3030... 3033... 3036... 3039... 3042... 3045... 3048... 3051... 3054... 3057... 3060... 3063... 3066... 3069... 3072... 3075... 3078... 3081... 3084... 3087... 3090... 3093... 3096... 3099... 3102... 3105... 3108... 3111... 3114... 3117... 3120... 3123... 3126... 3129... 3132... 3135... 3138... 3141... 3144... 3147... 3150... 3153... 3156... 3159... 3162... 3165... 3168... 3171... 3174... 3177... 3180... 3183... 3186... 3189... 3192... 3195... 3198... 3201... 3204... 3207... 3210... 3213... 3216... 3219... 3222... 3225... 3228... 3231... 3234... 3237... 3240... 3243... 3246... 3249... 3252... 3255... 3258... 3261... 3264... 3267... 3270... 3273... 3276... 3279... 3282... 3285... 3288... 3291... 3294... 3297... 3300... 3303... 3306... 3309... 3312... 3315... 3318... 3321... 3324... 3327... 3330... 3333... 3336... 3339... 3342... 3345... 3348... 3351... 3354... 3357... 3360... 3363... 3366... 3369... 3372... 3375... 3378... 3381... 3384... 3387... 3390... 3393... 3396... 3399... 3402... 3405... 3408... 3411... 3414... 3417... 3420... 3423... 3426... 3429... 3432... 3435... 3438... 3441... 3444... 3447... 3450... 3453... 3456... 3459... 3462... 3465... 3468... 3471... 3474... 3477... 3480... 3483... 3486... 3489... 3492... 3495... 3498... 3501... 3504... 3507... 3510... 3513... 3516... 3519... 3522... 3525... 3528... 3531... 3534... 3537... 3540... 3543... 3546... 3549... 3552... 3555... 3558... 3561... 3564... 3567... 3570... 3573... 3576... 3579... 3582... 3585... 3588... 3591... 3594... 3597... 3600... 3603... 3606... 3609... 3612... 3615... 3618... 3621... 3624... 3627... 3630... 3633... 3636... 3639... 3642... 3645... 3648... 3651... 3654... 3657... 3660... 3663... 3666... 3669... 3672... 3675... 3678... 3681... 3684... 3687... 3690... 3693... 3696... 3699... 3702... 3705... 3708... 3711... 3714... 3717... 3720... 3723... 3726... 3729... 3732... 3735... 3738... 3741... 3744... 3747... 3750... 3753... 3756... 3759... 3762... 3765... 3768... 3771... 3774... 3777... 3780... 3783... 3786... 3789... 3792... 3795... 3798... 3801... 3804... 3807... 3810... 3813... 3816... 3819... 3822... 3825... 3828... 3831... 3834... 3837... 3840... 3843... 3846... 3849... 3852... 3855... 3858... 3861... 3864... 3867... 3870... 3873... 3876... 3879... 3882... 3885... 3888... 3891... 3894... 3897... 3900... 3903... 3906... 3909... 3912... 3915... 3918... 3921... 3924... 3927... 3930... 3933... 3936... 3939... 3942... 3945... 3948... 3951... 3954... 3957... 3960... 3963... 3966... 3969... 3972... 3975... 3978... 3981... 3984... 3987... 3990... 3993... 3996... 3999... 4002... 4005... 4008... 4011... 4014... 4017... 4020... 4023... 4026... 4029... 4032... 4035... 4038... 4041... 4044... 4047... 4050... 4053... 4056... 4059... 4062... 4065... 4068... 4071... 4074... 4077... 4080... 4083... 4086... 4089... 4092... 4095... 4098... 4101... 4104... 4107... 4110... 4113... 4116... 4119... 4122... 4125... 4128... 4131... 4134... 4137... 4140... 4143... 4146... 4149... 4152... 4155... 4158... 4161... 4164... 4167... 4170... 4173... 4176... 4179... 4182... 4185... 4188... 4191... 4194... 4197... 4200... 4203... 4206... 4209... 4212... 4215... 4218... 4221... 4224... 4227... 4230... 4233... 4236... 4239... 4242... 4245... 4248... 4251... 4254... 4257... 4260... 4263... 4266... 4269... 4272... 4275... 4278... 4281... 4284... 4287... 4290... 4293... 4296... 4299... 4302... 4305... 4308... 4311... 4314... 4317... 4320... 4323... 4326... 4329... 4332... 4335... 4338... 4341... 4344... 4347... 4350... 4353... 4356... 4359... 4362... 4365... 4368... 4371... 4374... 4377... 4380... 4383... 4386... 4389... 4392... 4395... 4398... 4401... 4404... 4407... 4410... 4413... 4416... 4419... 4422... 4425... 4428... 4431... 4434... 4437... 4440... 4443... 4446... 4449... 4452... 4455... 4458... 4461... 4464... 4467... 4470... 4473... 4476... 4479... 4482... 4485... 4488... 4491... 4494... 4497... 4500... 4503... 4506... 4509... 4512... 4515... 4518... 4521... 4524... 4527... 4530... 4533... 4536... 4539... 4542... 4545... 4548... 4551... 4554... 4557... 4560... 4563... 4566... 4569... 4572... 4575... 4578... 4581... 4584... 4587... 4590... 4593... 4596... 4599... 4602... 4605... 4608... 4611... 4614... 4617... 4620... 4623... 4626... 4629... 4632... 4635... 4638... 4641... 4644... 4647... 4650... 4653... 4656... 4659... 4662... 4665... 4668... 4671... 4674... 4677... 4680... 4683... 4686... 4689... 4692... 4695... 4698... 4701... 4704... 4707... 4710... 4713... 4716... 4719... 4722... 4725... 4728... 4731... 4734... 4737... 4740... 4743... 4746... 4749... 4752... 4755... 4758... 4761... 4764... 4767... 4770... 4773... 4776... 4779... 4782... 4785... 4788... 4791... 4794... 4797... 4800... 4803... 4806... 4809... 4812... 4815... 4818... 4821... 4824... 4827... 4830... 4833... 4836... 4839... 4842... 4845... 4848... 4851... 4854... 4857... 4860... 4863... 4866... 4869... 4872... 4875... 4878... 4881... 4884... 4887... 4890... 4893... 4896... 4899... 4902... 4905... 4908... 4911... 4914... 4917... 4920... 4923... 4926... 4929... 4932... 4935... 4938... 4941... 4944... 4947... 4950... 4953... 4956... 4959... 4962... 4965... 4968... 4971... 4974... 4977... 4980... 4983... 4986... 4989... 4992... 4995... 4998... 5001... 5004... 5007... 5010... 5013... 5016... 5019... 5022... 5025... 5028... 5031... 5034... 5037... 5040... 5043... 5046... 5049... 5052... 5055... 5058... 5061... 5064... 5067... 5070... 5073... 5076... 5079... 5082... 5085... 5088... 5091... 5094... 5097... 5100... 5103... 5106... 5109... 5112... 5115... 5118... 5121... 5124... 5127... 5130... 5133... 5136... 5139... 5142... 5145... 5148... 5151... 5154... 5157... 5160... 5163... 5166... 5169... 5172... 5175... 5178... 5181... 5184... 5187... 5190... 5193... 5196... 5199... 5202... 5205... 5208... 5211... 5214... 5217... 5220... 5223... 5226... 5229... 5232... 5235... 5238... 5241... 5244... 5247... 5250... 5253... 5256... 5259... 5262... 5265... 5268... 5271... 5274... 5277... 5280... 5283... 5286... 5289... 5292... 5295... 5298... 5301... 5304... 5307... 5310... 5313... 5316... 5319... 5322... 5325... 5328... 5331... 5334... 5337... 5340... 5343... 5346... 5349... 5352... 5355... 5358... 5361... 5364... 5367... 5370... 5373... 5376... 5379... 5382... 5385... 5388... 5391... 5394... 5397... 5400... 5403... 5406... 5409... 5412... 5415... 5418... 5421... 5424... 5427... 5430... 5433... 5436... 5439... 5442... 5445... 5448... 5451... 5454... 5457... 5460... 5463... 5466... 5469... 5472... 5475... 5478... 5481... 5484... 5487... 5490... 5493... 5496... 5499... 5502... 5505... 5508... 5511... 5514... 5517... 5520... 5523... 5526... 5529... 5532... 5535... 5538... 5541... 5544... 5547... 5550... 5553... 5556... 5559... 5562... 5565... 5568... 5571... 5574... 5577... 5580... 5583... 5586... 5589... 5592... 5595... 5598... 5601... 5604... 5607... 5610... 5613... 5616... 5619... 5622... 5625... 5628... 5631... 5634... 5637... 5640... 5643... 5646... 5649... 5652... 5655... 5658... 5661... 5664... 5667... 5670... 5673... 5676... 5679... 5682... 5685... 5688... 5691... 5694... 5697... 5700... 5703... 5706... 5709... 5712... 5715... 5718... 5721... 5724... 5727... 5730... 5733... 5736... 5739... 5742... 574

Tree alignment

- Ideally:
 - Find alignment that maximizes probability that sequences evolved from common ancestor



Multiple sequence alignment

Bioinformatics Fall 2011

7

Tree alignment

- Model the k sequences with a tree having k leaves (1 to 1 correspondence)
- Assign sequences to internal nodes
- Compute a weight for each edge, which is the similarity score
- Sum of all the weights is the score of the tree
- Choose internal nodes in order to maximize the scores.

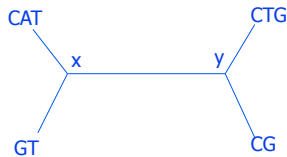
Multiple sequence alignment

Bioinformatics Fall 2011

8

Tree alignment example

- Match +1, gap -1, mismatch 0



- If $x=CT$ and $y=CG$, score = 6

Multiple sequence alignment

Bioinformatics Fall 2011

9

Analysis

- The tree alignment problem is NP-complete
 - Holds even for the special case of star alignment
 - “lifting alignment” gives a 2-approximate algorithm
 - Assign one of the leaves to the parent
 - Choose over all such assignments
- The generalized tree alignment problem (find the best tree) is also NP-complete

Multiple sequence alignment

Bioinformatics Fall 2011

10

Consensus representations

- Relative frequencies of symbols in each column
 - Adds up to 1 in each column (PSSM)
- Steiner string
 - Minimize the error with a center string (*Steiner error*)
 - May not belong to the set of input strings
- Consensus string for a given multiple alignment
 - Choose optimal character in every column
 - Consensus string is the concatenation of these characters
 - Alignment error of a column is the distance-sum to the optimal character of all symbols in the column
 - Alignment error* of a consensus string is the sum of all column errors
- Optimal consensus string: optimize over all multiple alignments
 - An optimal consensus string is also a Steiner string and vice versa.
- Signature representation
 - Regular expression
 - Helicase protein: $[&H][&A]D[DE]x_n[TSN]x_n[QK]Gx_r[&A]$
 - x & r is any amino acid in $\{L,L,V,M,F,Y,W\}$

Multiple sequence alignment

Bioinformatics Fall 2011

11

Steiner string and consensus error metric

- Minimize $\sum D(s, x_i)$, over all possible strings s
 - Called the *Steiner error*
- String s_{\min} is called the *Steiner string*
 - May not belong to the set of inputs
 - NP-complete
- Center string* provides an approximation factor of 2
 - Proved later
- Steiner string* provides the optimal consensus string

Multiple sequence alignment

Bioinformatics Fall 2011

12

Scoring Function: Sum of Pairs

Definition: Induced pairwise alignment

A pairwise alignment induced by the multiple alignment

Example:

x: AC-GCGG-C
y: AC-GC-GAG
z: GCCGC-GAG

Induces:

x: ACGCGG-C; x: AC-GCGG-C; y: AC-GCGAG
y: ACGC-GAC; z: GCCGC-GAG; z: GCCGCGAG

Multiple sequence alignment

Bioinformatics Fall 2011

13

Sum of Pairs (cont' d)

- The *sum-of-pairs* (SP) score of a multiple alignment A is the sum of the scores of all induced pairwise alignments

$$S(A) = \sum_{i < j} S(A_{ij})$$

A_{ij} is the induced alignment of x_i, x_j

- Drawback: no evolutionary characterization
 - Every sequence derived from all others

Multiple sequence alignment

Bioinformatics Fall 2011

14

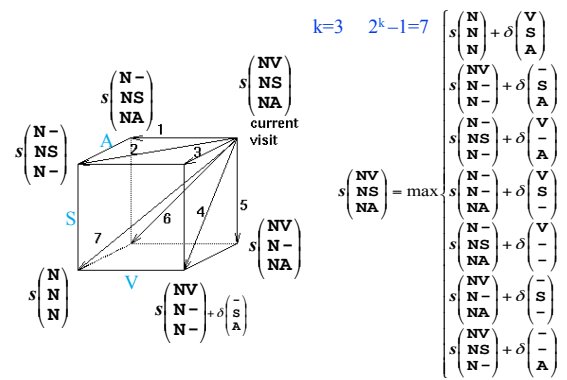
Optimal solution for SP scores

- Multidimensional Dynamic Programming
- Generalization of pair-wise alignment
- For simplicity, assume k sequences of length n
- The dynamic programming array is k -dimensional hyperlattice of length $n+1$ (including initial gaps)
- The entry $F(i_1, \dots, i_k)$ represents score of optimal alignment for $s_1[1..i_1], \dots, s_k[1..i_k]$
- Initialize values on the faces of the hyperlattice

Multiple sequence alignment

Bioinformatics Fall 2011

15



Multiple sequence alignment

Bioinformatics Fall 2011

16

Complexity

- Space complexity: $O(n^k)$ for k sequences, each n long.
- Computing at a cell: $O(2^k)$. cost of computing δ .
- Time complexity: $O(2^k n^k)$. cost of computing δ .
- Finding the optimal solution is exponential in k
- Proven to be NP-complete for a number of cost functions

Multiple sequence alignment

Bioinformatics Fall 2011

17

Faster Dynamic Programming

- Carrillo and Lipman 88 (CL)
- Pruning of hyperlattice in DP
- Practical for about 6 sequences of length about 200.

Multiple sequence alignment

Bioinformatics Fall 2011

18

Star alignment

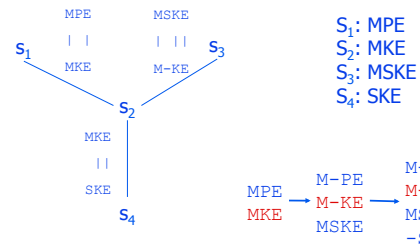
- Heuristic for multiple sequence alignments
- Select a sequence c as the center of the star
- For each sequence x_1, \dots, x_k such that index $i \neq c$, perform a Needleman-Wunsch global alignment
- Aggregate alignments with the principle "once a gap, always a gap."

Multiple sequence alignment

Bioinformatics Fall 2011

19

Star alignment example



Multiple sequence alignment

Bioinformatics Fall 2011

20

Choosing a center

- Let $D(x_i, x_j)$ be the optimal distance between sequences x_i and x_j .
- Given a multiple alignment A , let $c(A_{ij})$ be the distance between x_i and x_j that is induced on account of A .
- Calculate all $O(k^2)$ alignments, and pick the sequence x_i that minimizes the following as the center string x_c

$$\sum_j D(x_i, x_j)$$
- The resulting multiple alignment A has the property that $c(A_{ci}) = D(x_c, x_i)$.

Multiple sequence alignment

Bioinformatics Fall 2011

21

Analysis

- Assume all sequences have length n
- $O(k^2 n^2)$ to calculate center
- Step i of iterative pairwise alignment takes $O(i \cdot n)$ time
 - Only align wrt the center string
 - Insert gaps into center string appropriately
 - Other strings index into the center string positions
- $O(k^2 n^2)$ overall cost
- Produces multiple sequence alignments whose SP (distance) values are at most twice that of the optimal solution, *provided triangle inequality holds*.

Multiple sequence alignment

Bioinformatics Fall 2011

22

Bound analysis

- Let $M = \sum_{i=2}^k c(A_{1i}) = \sum_{i=2}^k D(x_1, x_i)$, assume x_1 is the center
- $2 c(A) = \sum_i \sum_{j \neq i} c(A_{ij}) \leq \sum_i \sum_{j \neq i} [c(A_{1i}) + c(A_{1j})] =$

$$2(k-1) \sum_{i=2}^k c(A_{1i}) = 2(k-1) M$$
- $2 c(A^*) = \sum_i \sum_{j \neq i} c(A^*_{ij}) \geq \sum_i \sum_{j \neq i} D(x_i, x_j) \geq k \sum_{i=2}^k D(x_1, x_i)$

$$= k M$$
- $c(A)/c(A^*) \leq 2(k-1)/k \leq 2$

Multiple sequence alignment

Bioinformatics Fall 2011

23

Steiner error

- Center string also provides an approximation factor of 2 under the steiner error metric
- Assume triangle inequality
- Let $E(x)$ denote the steiner error wrt string x .
- Let z be the Steiner string
 - $E(z) = \sum_i D(z, x_i)$

Multiple sequence alignment

Bioinformatics Fall 2011

24

Steiner error

- For any string y in the input set and steiner string z ,
 - $E(y) = \sum_i D(y, x_i) \leq \sum_{y \neq x_i} [D(y, z) + D(z, x_i)] = (k-2) D(y, z) + D(y, z) + \sum_{y \neq x_i} D(z, x_i) = (k-2) D(y, z) + E(z)$
- Pick y^* from input set that is closest to z .
 - $E(z) = \sum_i D(z, x_i) \geq k D(y^*, z)$
- $E(y^*)/E(z) \leq [(k-2) D(y^*, z) + E(z)]/E(z) \leq (k-2) D(y^*, z) / [k D(y^*, z)] + 1 \leq 2-2/k \leq 2$
- $E(c) \leq E(y^*)$, c is the center string
- $E(c)/E(z) \leq 2$

ClustalW

- Progressive alignment
- 3 steps:
 - All pairs of sequences are aligned to produce a distance matrix (or a similarity matrix)
 - A rooted guide tree is calculated from this matrix by the neighbor-joining (NJ) method
 - Neighbor Joining – Saitou, 1987
 - The sequences are aligned progressively according to the branching order in the guide tree

ClustalW example

S₁ ALSK
 S₂ TNSD
 S₃ NASK
 S₄ NTSD

ClustalW example

S₁ ALSK
 S₂ TNSD
 S₃ NASK
 S₄ NTSD

All pairwise alignments

	S ₁	S ₂	S ₃	S ₄
S ₁	0	9	4	7
S ₂		0	8	3
S ₃			0	7
S ₄				0

ClustalW example

S₁ ALSK
 S₂ TNSD
 S₃ NASK
 S₄ NTSD

All pairwise alignments

	S ₁	S ₂	S ₃	S ₄
S ₁	0	9	4	7
S ₂		0	8	3
S ₃			0	7
S ₄				0

Neighbor Joining

S₁
 S₃
 S₂
 S₄

ClustalW example

S₁ ALSK
 S₂ TNSD
 S₃ NASK
 S₄ NTSD

Multiple Alignment Steps

- Align S₁ with S₃
- Align S₂ with S₄
- Align (S₁, S₃) with (S₂, S₄)

All pairwise alignments

	S ₁	S ₂	S ₃	S ₄
S ₁	0	9	4	7
S ₂		0	8	3
S ₃			0	7
S ₄				0

Neighbor Joining

S₁
 S₃
 S₂
 S₄

ClustalW example

All pairwise alignments

	S ₁	S ₂	S ₃	S ₄
S ₁	0	9	4	7
S ₂		0	8	3
S ₃			0	7
S ₄				0

Distance Matrix
Multiple sequence alignment

Multiple Alignment Steps

1. Align S₁ with S₃
2. Align S₂ with S₄
3. Align (S₁, S₃) with (S₂, S₄)

How to align profiles?

Neighbor Joining

Rooted Tree

S₁ ALSK
-ALSK
-ALSK

S₂ TNSD
-TNSD
NA-SK

S₃ NASK
NA-SK
-TNSD

S₄ NTSD
NT-SD
NT-SD

Bioinformatics Fall 2011 **31**

Other progressive approaches

- PILEUP
 - Similar to CLUSTALW
 - Uses UPGMA to produce tree

Bioinformatics Fall 2011 **32**

Problems with progressive alignments

- Depend on pairwise alignments
- If sequences are very distantly related, much higher likelihood of errors
- Frozen initial alignments

Bioinformatics Fall 2011 **33**

Frozen initial alignment

- Initial alignments are “frozen” even when new evidence comes

Example:

```

x: GAAGTT
y: GAC-TT > Frozen!

z: GAACTG
w: GTACTG > Now clear that correct y = GA-CTT
  
```

Bioinformatics Fall 2011 **34**

Profile HMM for multiple alignment

- Train a profile HMM with the given sequences
 - Baum Welch or Viterbi training
- Find the most probable path for each sequence
- Use the corresponding states for constructing a multiple alignment

Bioinformatics Fall 2011 **35**

Multiple alignment tools

- Clustal W (Thompson, 1994)
 - Most popular
- PRRP (Gotoh, 1993)
- HMMT (Eddy, 1995)
- DIALIGN (Morgenstern, 1998)
- T-Coffee (Notredame, 2000)
- MUSCLE (Edgar, 2004)
- Align-m (Walle, 2004)
- PROBCONS (Do, 2004)

Bioinformatics Fall 2011 **36**

Evaluating multiple alignments

- Balibase benchmark (Thompson, 1999)
- De facto standard for assessing the quality of a multiple alignment tool
- Manually refined multiple sequence alignments
- Quality measured by how good it matches the core blocks

Multiple sequence alignment

Bioinformatics Fall 2011

37

Quick Primer on NP completeness

- Polynomial-time reductions
 - If we could solve X in polynomial time, then we could also solve Y in polynomial time
 - $Y \leq_p X$
- Class NP
 - Set of all problems for which there exists an efficient certifier
- $P = NP?$
 - *checking a solution versus finding a solution*

Multiple sequence alignment

Bioinformatics Fall 2011

38

- NP-completeness
 - X is NP-complete if
 - $X \in NP$
 - For all $Y \in NP$, $Y \leq_p X$
 - If X is NP-complete, then X is solvable in polynomial time iff $P=NP$
 - Satisfiability is NP-complete
 - If Y is NP-complete and X is in NP with the property that $Y \leq_p X$, then X is NP complete

Multiple sequence alignment

Bioinformatics Fall 2011

39