Since their discovery in 1928, antibiotics have played a key role in the treatment and prevention of bacterial infections. Antibiotics are a class of molecules that can inhibit the growth of or destroy bacteria or other microorganisms. They are largely produced by organisms such as fungi to kill or otherwise ward off competing or harmful microorganisms. To put it simply, antibiotics have the capacity to significantly diminish microorganism presence in the human body while minimally harming said individual. When harvested for medicinal use, antibiotics have the unparalleled capacity to treat infections that could prove hazardous to human health.

While previously discovered antibiotics have proven invaluable in medicinal practice, a growing number of superbugs resistant to these antibiotics are greatly diminishing their effectiveness. As such, the development of new antibiotics is a critical challenge in modern medicine. “A difficult problem in antibiotics research is that of sequencing newly discovered antibiotics, or determining the order of amino acids making up the antibiotic peptide.”

The composition of most antibiotics varies greatly from the sequences of known typical polypeptides. There are three main differences. First, many antibiotics are circular peptides, not linear. Without a specific beginning or end, these “cyclopeptides” can be difficult to sequence. Second, organisms that produce cyclopeptides do not always do so through the means described by the Central Dogma of Molecular Biology, which states that DNA makes RNA which makes proteins. And lastly, because these cyclopeptides are created through a mechanism independent of the Central Dogma, additional amino acids can be incorporated into their structure that are not included in the 20 proteinogenic amino acids. This increases the number of viable amino acids to something closer to 100.

These differences make sequencing cyclopeptides something of a challenge to bioinformaticians. Normally polypeptide compositions are determined via reliance on databases of known polypeptides or on DNA sequences that coincide with the polypeptide. Because very few cyclopeptides have been sequenced and because they don’t directly rely on DNA to be produced, one must rely on *de novo* sequencing techniques to determine their sequence.

Most *de novo* cyclopeptide sequencing techniques rely on expensive machinery to function. The ideal scenario would be to sequence a cyclopeptide from data produced by a mass spectrometer. A mass spectrometer shatters molecules into pieces and then weighs the resulting fragments in daltons (Da). In sequencing cyclopeptides, one would shatter several copies of a cyclopeptide and use the resulting fragments to reconstruct the cyclopeptide. Unfortunately, mass spectrometry data is both incomplete and impure. Something close to two thirds of data is false data, and of the third that is accurate only a portion of the total possible cyclopeptide fragments are represented.

Here is an example. Tyrocidine B1 is an antibiotic produced by *Bacillus brevis*. It is composed of the following 10 amino acids:

Val\* – Lys – Leu – Phe – Pro – Trp – Phe – Asn – Gln – Tyr

\*Note that, because this is a cyclopeptide, Valine is not actually the “start” of the peptide. The sequence could have started at Phe, wrapped around to Leucine, and would have been just as viable of a sequence.

There are several fragments that could show up in a mass spectrometer run of Tyrocidine B1. For example, the amino acid weight of Valine (99.06) could appear. The weight of Val-Lys-Leu (99.06+128.09+113.09 = 340.24) could also appear. The following is a list of all valid fragments that COULD show up in a mass spectrometer run of Tyrocidine B1 (the “Theoretical Spectrum”):

97.05276, 99.06841, 113.08406, 114.04293, 128.09496, 147.06841, 163.06333, 186.07931, 227.16337, 241.17901999999998, 242.13788999999997, 244.12117, 260.15247, 261.11134, 262.13174000000004, 283.13207, 291.15828999999997, 333.14772, 340.24743, 357.20523000000003, 388.24743, 389.20629999999994, 390.22669999999994, 390.22670000000005, 405.20122, 430.20047999999997, 447.19065, 485.30019000000004, 487.31584, 503.3107600000001, 504.26963, 518.32166, 543.28454, 544.2434099999999, 552.26963, 575.28561, 577.2688899999999, 584.3686, 631.40572, 632.36459, 650.37917, 651.33804, 671.3795, 672.3383699999999, 690.35295, 691.3118199999999, 738.3489400000001, 745.44865, 747.4319300000001, 770.44791, 778.47413, 779.433, 804.3958799999999, 818.44791, 819.4067799999999, 835.4016999999999, 837.41735, 875.52689, 892.51706, 917.51632, 932.4908399999999, 933.51124, 934.4701099999999, 965.5123100000001, 982.47011, 989.56982, 1031.55925, 1039.58547, 1060.5857999999998, 1061.6062, 1062.5650699999999, 1078.59637, 1080.5796500000001, 1081.53852, 1095.5541699999999, 1136.63823, 1159.65421, 1175.6491299999998, 1175.64913, 1194.62258, 1194.6225800000002, 1208.67461, 1209.63348, 1223.6491299999998, 1225.66478, 1322.7

The following is an example of an actual Experimental Spectrum of Tyrosidine B1, with datapoints within 0.3 daltons of a data point from the Theoretical Spectrum in bold. Note that all data points have a +1 charge as a technical side effect of mass spectrometry.

372.2, 397.2, 402.0, **406.3**, 415.1, **431.2**, **448.3**, 449.3, 452.2, 471.3, **486.3**, **488.2**, 500.5, **505.3**, 516.1, 536.1, **544.2**, **545.3**, 562.5, 571.3, 599.2, 614.4, 615.4, 616.4, 618.2, **632.0**, 655.5, 656.3, **672.5**, **673.3**, 677.3, **691.4**, **692.4**, 712.1, 722.3, **746.5**, 760.4, 761.6, 762.5, **771.6**, 788.4, 802.3, 803.3, 818.5, **819.4**, 831.4, **836.3**, 853.3, 875.5, **876.5**, 901.5, 915.9, 916.5, 917.8, **918.4**, **933.4,** **934.7**, **935.5**, 949.4, **966.2**, 995.4, 1015.6, 1027.5, 1029.5, 1031.5, 1044.5, 1046.5, **1061.5**, **1063.4**, **1079.2**, 1083.7, 1088.4, 1093.5, **1096.5**, 1098.4, 1158.5, 1159.5, **1176.6**, 1177.7, 1178.6, 1192.7, **1195.4**, 1207.5, **1210.4**, **1224.6**, 1252.5, 1270.5, 1271.5, 1278.6, 1279.6, 1295.6, 1305.6, 1306.5, 1307.5, 1309.6

The challenge is: Can you reconstruct Tyrosidine B1 from the Experimental Spectrum given above (without knowing in advance which ones correspond to the Theoretical Spectrum)? That is the present challenge of sequencing cyclopeptides.

All data and concepts taken from the book Bioinformatics Algorithms: An Active Learning Approach, chapter 4.