Abstract

Proteins are large, complex molecules that play many vital roles in the body. It is known that proteins usually don't work individually but team up forming assemblies known as protein complexes. Therefore, there is a lot of interest in inferring that a group of proteins work together as a complex. Quantitative proteomics data from mass spectrometry experiments measures protein abundances across multiple conditions such as cell types providing expression profiles for the proteins in question. Similar profiles indicate that the proteins act in concert and form a complex. In this work, we propose a computational approach to verifying complex composition using quantitative proteomics data from the human proteome, provided in the ProteomicsDB repository. Our new *in-silico* approach investigates the following questions for a given set of proteins: (i) if an individual protein is part of the potential complex, (ii) and if the entire set acts synergistically as a complex. For an in-depth assessment of complex composition, we implemented four different statistical models, and used two different randomization approaches to assess statistical significance. To validate our statistical models, we tested them on complex compositions data extracted from CORUM and the Complex Portal. Our results, for significance level $\alpha = 0.1$, show that our tool could successfully verify 73% and 66% of the complexes with at least two proteins, and 89% and 87% of the complexes with at least five proteins, on CORUM and Complex Portal respectively. Therefore, we provide a computational tool that can be used to complement or even substitute expensive laboratory experiments to assess protein complex composition.

Proteiner er store, komplekse molekyler, som har mange essentielle roller i kroppen. Man ved at proteiner som regel ikke arbejder individuelt, men finder sammen i proteinkomplekser. Derfor er der meget der tyder på, at en gruppe af proteiner arbejder sammen som et kompleks. Kvantitative proteomisk data fra massespektrometri eksperimenter måler hyppighed af proteiner over forskellige betingelser, så som ekspressionsprofiler af proteiner fra forskellige celletyper. Lignende profiler indikerer at proteiner arbejder sammen og danner kompleks. I dette projekt, foreslår vi en beregningsmæssig fremgangsmåde til at verificere sammensætninger af komplekser ved at bruge kvantitativ proteomik data fra menneskets proteom som kan hentes fra ProteomicsDB repository. Vores nye in-silico fremgangsmåde undersøger følgende spørgsmål for en givet mængde af proteiner: (i) hvis et individuelt protein er en del af det potentielle kompleks, (ii) og hvis hele mængden af proteiner arbejder synergisk som et kompleks. For en dybdegående vurdering af komplekssammensætning, har vi implementeret fire forskellige statistiske modeller og brugt to forskellige tilfældighedsfremgangsmåder til at vurdere statistisk signifikans. For at validere vores statistisk, har vi testet dem under et komplekst sammensat data ekstraheret fra CORUM og Complex Portal. Vores resultater med signifikans niveau på $\alpha = 0.1$, viser at vores værktøj successfuldt verificere 73% og 66% af komplekserne for mindst to proteiner, og 89% og 87% af komplekserne for mindst fem proteiner på CORUM og Complex Portal. Vi leverer derfor et beregningsmæssigt værktøj, der kan bruges til at komplementere og tilmed også erstatte dyre laboratorieeksperimenter til at vurdere sammensætningen af proteinkomplekser.