UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE QUÍMICA E BIOQUÍMICA

How to Disassemble a Virus Capsid: A Computational Approach

Claudio Alexandre Piedade

DISSERTAÇÃO
MESTRADO EM BIOQUÍMICA
ESPECIALIDADE EM BIOQUÍMICA

Dissertação orientada por:
Prof. António Ferreira

2016

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE QUÍMICA E BIOQUÍMICA

How to Disassemble a Virus Capsid: A Computational Approach

Claudio Alexandre Piedade

DISSERTAÇÃO
MESTRADO EM BIOQUÍMICA
ESPECIALIDADE EM BIOQUÍMICA

Dissertação orientada por:
Prof. António Ferreira

2016

An expert problem solver must be endowed with two incompatible qualities – a restless imagination and a patient pertinacity.

– Howard W. Eves

Abstract

Viruses are one of the main subjects of study in science due to the amount of diseases and deaths they cause, not only in humans, but also in other organisms, such as plants, other mammals, insects and microorganisms. This raises the need to understand the mechanisms of host infection. Capsids surround the genetic information of viruses and many experimental and theoretical studies have been done to study the icosahedral virus capsid assembly, but very few have focused on the disassembly process. These rare studies point to the loss of a triangle-shaped trimer as the first step of disassembly, or to the loss of a pentagon-shaped pentadecamer (15-mer). Also, no study was found that approached the problem of disassembling virus through a 60-subunit model using combinatorial geometry and symmetry groups.

In this work, we tried to predict the sequence of subunit loss from the virus capsids using a combinatorial and geometric approach, taking into account the equivalence of subunit configuration and using cyclical permutations of the Deltoidal Hexecontahedron symmetry group, applied on a binary vector. The aim of this study was to assess whether there was a similar process of capsid disassembly throughout different viruses' families and, if not, try to find a pattern of disassembly characteristic of a given family. Using $T = 1$ viruses' capsids, we analysed 51 known structures that were subdivided into 14 groups and predicted the sequence of the loss of up to 5 subunits. The energy of the different capsid configurations was estimated by three different heuristics, all based only on the number of weak inter-subunit contacts.

Our results show a prevalence on the lost of a triangle-shaped trimer of subunits in several of the groups identified, mainly those comprising Parvoviruses, Circoviruses, Birnaviruses and Hepatitis E Viruses. Densoviruses (a parvovirus) also, in general, loose a square-shaped tetramer

while Human Adenoviruses Pt-Dd loose a pentagon-shaped pentamer. All these viruses, except for Human Adenoviruses Pt-Dd, loose a right-trapezium or a scalene-trapezium-shaped pentamer, built by the lost of a triangular trimer and two adjacent subunits. This work shows that there is a main disassembly pathway, through the removal of a triangle trimer, although it can not be generalized to all icosahedral viruses. These predictions have a very low dependency on the heuristical measure of total energy used. This triangular trimer might be an interesting target for novel antiviral drugs, by interfering with the capsid disassembly.

Keywords: virus, capsid disassembly, computational biology, combinatorial geometry, symmetry groups, structural biology

Resumo

Os vírus são um dos mais importantes objectos de estudo em ciência devido ao elevado número de doenças e mortes que os mesmos causam, não só em humanos, mas também noutros organismos, tais como plantas, outros mamíferos, insectos ou mesmo microrganismos. Vírus como o da gripe causam por ano entre 250 a 500 milhões de mortes todos os anos. Por outro lado, o Vírus da Imunodeficiência Humana afecta certa de 0.8% da população mundial, com idades compreendidas entre os 15 e 49 anos, causando por ano 1.1 milhões de mortes. A constante presença de vírus no nosso dia-a-dia leva-nos a investigar sobre como combatê-los, bem como a estudarmos os seus modos de infecção dos hospedeiros.

Os vírus são estruturas nucleoproteícas, compostas por material genético e proteínas envolventes, constituindo a cápside. Alguns vírus apresentam uma bicamada lipídica com proteínas e/ou glicoproteínas de hospedeiros anteriores, de modo a facilitar a interacção e ligação à superfície das células dos novos hospedeiros, bem como a sua infecção. As cápsides podem apresentar diferentes organizações espaciais das suas proteínas, podendo ser organizadas de forma helicoidal, redonda ou complexa. A forma de organizar a cápside de um modo redondo está associado a uma simetria icosaédrica, sendo as subunidades proteicas dispostas ao longo das faces triangulares do icosaedro. É de notar que as cápsides icosaédricas mais simples, onde não há triangulação das faces do icosaedro ($T = 1$) são compostas por 60 subunidades, havendo três proteínas por cada uma das 20 faces do icosaedro.

Para entender a infecção dos hospedeiros, é importante saber como a cápside se desmonta, permitindo o acesso do material genético à maquinaria celular. Alguns estudos experimentais e teóricos foram realizados com o intuito de observar ou prever o modo como as cápsides se montam, mas

poucos estudos existem sobre o processo opostos, o de desmontagem. Um dos estudos mais recentes indica que o primeiro passo para a desmontagem decorre pela remoção de um trímero disposto numa conformação triangular ou pela remoção de um pentadecamero (15 subunidades) com 5 triângulos dispostos de modo pentagonal, cada triângulo com um trímero. Em trabalhos anteriores, são usados modelos em que o número de subunidades é reduzido para aumentar a capacidade de computação das possíveis combinações, modelos esses baseados na geometria do icosaedro ou do dodecaedro. Para além disso, nenhum estudo encontrado trata a desmontagem da cápside viral com modelos de 60 subunidades, nem com bases matemáticas como a geometria, combinatória e simetria.

Neste trabalho, abordou-se a remoção de subunidades de uma cápside viral de uma perspectiva combinatória, tendo em conta a geometria e a simetria virais, fazendo uso, como poliedro de referência, o Hexecontaedro Deltoidal. Para tal, construiu-se um grupo de simetria para cápsides virais $T = 1$, de modo a reduzir o peso combinatorial, comparando se duas estruturas eram equivalentes, recorrendo à simetria. De modo a confirmar se o número de combinações estava correcto, usou-se o Lema de Contagem de Burnside.

Este estudo teve como objectivo estabelecer se existia um processo de desmontagem de cápsides comum em diferentes famílias de vírus e, caso assim não fosse, encontrar um processo de desmontagem característico de cada família de vírus.

De um modo computacional, previu-se a sequência de remoção de subunidades em diferentes cápsides virais. Restringindo o estudo a cápsides virais $T = 1$, foram analisadas 51 estruturas, subdivididas em 14 grupos, sendo, posteriormente e iterativamente, removidas até 5 subunidades das suas cápsides. Para cada estrutura possível, foi analisada se existiam configurações equivalentes (e redundantes), utilizando o grupo de simetria. Para cada combinação possível, foram aplicadas as diferentes operações de simetria, representada como uma permutação cíclica de um vector binário (ausente ou presente), onde foram ignoradas combinações que

gerassem redundância. Para além disto, o uso da representação das subunidades como forma de um grafo das faces do Hexecontaedro Deltoidal foi essencial para o estudo da formação de subgrafos desconectados do grafo principal, sendo removidos os subgrafos de menor dimensão.

Outro foco do trabalho foi a comparação de diferentes heurísticas para medir a energia dos diferentes fragmentos de cápside formados no processo de desmontagem. A conjugação de aspectos geométricos e combinatórios visam cobrir um objectivo a longo termo de ser possível prever o percurso de desmontagem de um modo computacional. Para cada estrutura não redundante foi calculada uma energia global da configuração de subunidades, através de uma medida heurística baseada exclusivamente no número de interacções entre subunidades. A cada combinação não redundante, foi criado um nodo num grafo em forma de árvore, em que os vértices unem as transições possíveis entre duas combinações, sendo estes vértices pesados pela variação da energia calculada pela heurística para transitar entre as duas combinações. Assim, foi calculado o trajecto com a menor variação de energia, desde a cápside total até à remoção de N subunidades. Todos os processos anteriormente descritos foram calculados utilizando a linguagem de programação Python versão 2.7, e estruturas das cápsides obtidas do *Protein Data Bank*.

Os resultados apresentados indicam uma prevalência na perda de um trímero de subunidades, com conformação triangular, num grande número dos grupos estudados, tais como os Parvovírus (Vírus Adeno-Associados, Parvovírus Bovino, Humano e Suíno, Protoparvovírus de Roedores, Vírus da Panleucopenia Canina e Felina), Birnavírus das Aves, Vírus da Hepatite E e Circovírus Suíno. Os Densovírus estudados (de *Bombyx mori*, *Galleria mellonella* e *Penaeus stylirostris*), também parvovírus, perdem, principalmente, quatro subunidades da cápside com uma forma quadrada. Em todos os resultados indicados acima, uma estrutura semelhante a um trapézio (rectangular ou escaleno) era removida quando se retiravam cinco proteínas, sendo que estas estruturas são formadas por uma região onde

foi removido um trímero triangular. Os Adenovírus Humanos Pt-Dd apresentaram uma tendência distinta dos restantes resultados, não removendo trímeros triangulares, mas sim subunidades ao longo do eixo de simetria rotacional de ordem 5, sendo que a conformação mais estável quando se retiram 5 proteínas é um pentâmero com conformação pentangular. Isto mostra que nem todas as desmontagens da cápside têm que envolver a remoção de trímeros triangulares, como previsto na literatura. Mesmo assim, não se poderá considerar isto um factor exclusivo dos Adenovírus Humanos Pt-Dd, pois a amostra de vírus analisada foi reduzida e, tal como se viu nos resultados anteriores, vários grupos de vírus seguiram a mesma linha de desmontagem. O uso de um modelo baseado em 60 subunidades permitiu o estudo de casos como o dos Adenovírus Humanos Pt-Dd em que não eram removidos triângulos, mas sim pentágonos pentaméricos.

As diferentes heurísticas não produziram resultados distintos entre si, sendo quase independentes dos diferentes pesos dados às interacções entre subunidades. A única diferença a apontar encontra-se no uso da heurística dependente exclusivamente do número de contactos (heurística II), apresentando processos de desmontagem das cápsides virais lineares, com baixa variabilidade de alternativas a seguir. Os resultados do grupo dos Vírus Satélite do Mosaico do Tabaco não foram analisados em detalhe, visto que se observou uma divergência bastante grande entre as quatro estruturas que formavam este grupo, em todas as heurísticas.

Este trabalho é uma contribuição nova para o estudo do processo de desmontagem de uma cápside viral, pela sua representação como um problema matemático da área da geometria e combinatória, algo ignorado ou não desenvolvido em profundidade pelos estudos encontrados na literatura. A compreensão do modo como podemos interferir nas ligações que asseguram a união entre as proteínas do trímero triangular e a restante cápside será importante para o desenvolvimento de drogas antivirais que consigam aumentar a estabilidade destas interacções, impedindo a desmontagem da cápside, ou diminuir a sua estabilidade, com a consequente desmontagem prematura, ambas diminuindo a taxa de infecção do hospedeiro.

Palavras Chave: vírus, desmontagem viral, biologia computacional, combinatória geométrica, grupos de simetria, biologia estrutural

Acknowledgements

Due to everyone mentioned being Portuguese, this section will be written in Portuguese.

Em primeiro lugar, tenho que agradecer ao meu orientador, Professor António Ferreira, por ter acreditado nas minhas capacidades, por me ter guiado em toda a dissertação, pelos conselhos e liberdade que me deu durante todo o desenvolvimento da dissertação. Ao Professor Carlos Cordeiro, agradeço todas as conversas e discussões científicas em volta (ou não) da dissertação, e pelos seus comentários sobre como guiar a dissertação. A ambos agradeço terem-me dado a oportunidade de demonstrar o meu valor.

Agradeço também a todos os Professores de Bioquímica que tive pelo conhecimento que me transmitiram, e em particular à Professora Ana Ponces-Freire, por ser uma inspiração para todas as gerações vindoras de bioquímicos.

À Doutora Ana Guerra e à Doutora Catarina Lucas, agradeço o suporte emocional que criaram em mim durante todo este ano, conseguindo desenvolver e manter um ser humano no seu estado de maior estabilidade, capaz da escrita de uma dissertação. Ao Professor Pedro Silva pela inspiração que deu para a resolução de alguns dos problemas desta dissertação.

Quero agradecer ao Filipe Gomes e ao Tiago Brito pelo o seu contributo, principalmente na parte mais matemática da dissertação, pois estiveram lá para me ajudar e/ou esclarecer dúvidas existenciais. Ao João Moutinho, pela opinão crítica, e ao João Morais, pelo ajuda mais computacional, sem as quais esta dissertação não teria sido conseguida. A ambos tenho que

agradecer aquele apoio social, sempre importante para nunca levar uma pessoa ao desespero intelectual. Ao *Rince Na Lisboa* por ter sido o meu labo B, me ter feito amar algo como nada antes e me ter feito dançar e viver a vida com um sorriso.

Tenho a agradecer aos meus amigos do secundário: Samuel Coelho, Magda Paiva, Inês Figueiredo, João Caramujo e Rúben Dias; os quais nunca me abandonaram e que sempre me apoiaram, mesmo após nos termos dividido em direcções opostas.

À minha família, tenho que agradecer por me terem dado todas as oportunidades que uma pessoa pode ter. Aos meus avós, tios-avós, tia, primos e cunhado, tenho que vos agradecer todo o apoio, toda a motivação e toda a vitalidade que me forneceram para eu chegar onde estou. À minha irmã, todas as discussões ciêntificas e compreensão que não consigo encontrar em mais lado nenhum. Aos meus pais, por se terem esforçado a vida toda para garantir a melhor educação, com a maior paciência e apoio possível, proporcionando todo o amor e carinho para uma vida completa.

Por último, agradeço à Fundação para a Ciência e Tecnologia pelo financiamento para o projecto "The Architecture of Life" (RECI/BBB-BEP/0104/2012).

Contents

1 Introduction **1**

1.1 Virus 1

1.2 Capsid Structure 2

1.3 Virus Assembly and Disassembly Studies 3

1.4 Aims 6

2 Methods **8**

2.1 Combinatorics and Symmetry 8

2.2 Structures of Viruses' Capsids and Energy Calculation 10

2.2.1 Hydrogen Bonds 10

2.2.2 Salt Bridges 11

2.2.3 Hydrophobic Contacts 11

2.2.4 Energy Calculation 11

2.3 Symmetry Group Construction 12

2.4 Removal of Proteins and Graph Representation 14

2.5 Optimal Path for Disassembly 15

2.6 Graph Representation 16

2.7 Implementation 16

3 Results **18**

3.1 Paths of Disassembly 18

3.2 Structural Paths 21

3.3 Effect of the Energy Heuristics 30

4 Discussion **43**

5 Conclusion **47**

References **48**

A Mathematical Background **I**
A.1 Combinatorics and Groups I
A.1.1 Combinations I
A.1.1.1 Why do Combinations Fail in our Problem? II
A.1.2 Symmetry Groups II
A.1.3 Cyclical Permutations and Symmetry Elements III
A.2 Burnside's Lemma and its Application V

B Permutation Group of the *I* Symmetric Group on a Deltoidal Hexecontahedron **VIII**

C Supplementary Results using Heuristic I **XV**

D Supplementary Results using Heuristic II **XIX**

E Supplementary Results using Heuristic III **XXX**

List of Figures

1.1 Tobacco Mosaic Virus structure, with the nucleic acid in the middle surrounded by the capsid proteins, as drawn by Klug and Caspar in 1960 (14). 3

1.2 Canine Panleukopenia Virus Capsid. $T = 1$ structure, where the triangle represents the face of a icosahedron, and the pentagon represents the pentamers around a 5-fold axis. Structure PDB_{ID} $1c8h$. 4

1.3 Scheme of the Hepatitis B Virus capsid ($T = 4$) assembly, with the formation of a nucleus, followed by the elongation by adding multiple dimeric subunits to form a complete capsid. Adapted from (17). . . . 5

2.1 Deltoidal Hexecontahedron, from (25). The highlighted region represents a triangular face of an Icosahedron. 9

2.2 Possible arrangements of the proteins on the triangular facets of the viruses capsids in a deltoidal (A) and a trapezoid (B) fashion. 9

2.3 Numbered faces of the Deltoidal Hexecontahedron as the proteins of a $T = 1$ virus capsid as found on Protein Data Bank. Five, three and two-fold axes marked with a pentagon, triangle and ellipse, respectively. 13

2.4 Geometrical vertices of the Rhombicosidodecahedron represented as graph vertices, with edges connecting the vertices that are adjacent by geometrical edges. This graph is equivalent to the graph of the faces of the Deltoidal Hexecontahedron represented as vertices and connected with the four adjacent faces by edges. 14

3.1 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Adeno-Associated Virus group in each step of the removal of one protein, with heuristic I. 19
3.2 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Bovine and Human Parvovirus groups (A), and Porcine Parvovirus (B) group, in each step of the removal of one protein, with heuristic I. 20
3.3 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the *Bombyx mori* (A), *Galleria mellonella* (B), *Penaeus stylirostris* (C) Densoviruses groups, in each step of the removal of one protein, with heuristic I. 22
3.4 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Canine and Feline Panleukopenia Virus group in each step of the removal of one protein, with heuristic I. 23
3.5 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Rodent Protoparvovirus group in each step of the removal of one protein, with heuristic I. 24
3.6 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Porcine Circovirus group in each step of the removal of one protein, with heuristic I. 25
3.7 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Avian Birnavirus group in each step of the removal of one protein, with heuristic I. 26
3.8 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Hepatitis E Virus group in each step of the removal of one protein, with heuristic I. 27
3.9 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Human Adenovirus group in each step of the removal of one protein, with heuristic I. 28
3.10 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Satellite Tobacco Mosaic Virus group in each step of the removal of one protein, with heuristic I. . . . 29

3.11 Graphs of the Rhombicosadodecahedron (A), indicating the combinations in which the following subunits are removed: $\{1,10,23\}$ (B), $\{1,2,22,23\}$ (C), $\{1,2,10,23\}$ (D). 31
3.12 Graphs of the Rhombicosadodecahedron, indicating the combinations in which the following subunits are removed: $\{1,2,10,22,23\}$ (A), $\{1,2,6,10,23\}$ (B), $\{1,6,13,45,46\}$ (C), $\{1,6,25,26,53\}$ (D), $\{1,2,3,4,5\}$ (E). 32
3.13 Structure of 1s58 (A) by the removal of proteins $\{1,10,23\}$ (B, close up: D), with the respective proteins coloured white (C). Structure painted by protein, from blue (first protein) to red (60^{th} protein) in a rainbow pattern. 33
3.14 Structure of 1dnv (A) by the removal of proteins $\{1,2,22,23\}$ (B, close up: D), with the respective proteins coloured white (C). Structure painted by protein, from blue (first protein) to red (60^{th} protein) in a rainbow pattern. 34
3.15 Structure of 3raa (A) by the removal of proteins $\{1,2,10,23\}$ (B, close up: D), with the respective proteins coloured white (C). Structure painted by protein, from blue (first protein) to red (60^{th} protein) in a rainbow pattern. 35
3.16 Structure of 1s58 (A) by the removal of proteins $\{1,2,10,22,23\}$ (B, close up: D), with the respective proteins coloured white (C). Structure painted by protein, from blue (first protein) to red (60^{th} protein) in a rainbow pattern. 36
3.17 Structure of 3p0s (A) by the removal of proteins $\{1,2,6,10,23\}$ (B, close up: D), with the respective proteins coloured white (C). Structure painted by protein, from blue (first protein) to red (60^{th} protein) in a rainbow pattern. 37
3.18 Structure of 2g8g (A) by the removal of proteins $\{1,6,13,45,46\}$ (B, close up from another perspective: D), with the respective proteins coloured white (C). Structure painted by protein, from blue (first protein) to red (60^{th} protein) in a rainbow pattern. 38

3.19 Structure of 2g8g (A) by the removal of proteins $\{1,6,25,26,53\}$ (B, close up: D), with the respective proteins coloured white (C). Structure painted by protein, from blue (first protein) to red (60^{th} protein) in a rainbow pattern. 39
3.20 Structure of 4aqq (A) by the removal of proteins $\{1,2,3,4,5\}$ (B, close up: D), with the respective proteins coloured white (C). Structure painted by protein, from blue (first protein) to red (60^{th} protein) in a rainbow pattern. 40

A.1 Vertex-Numbered Square C_4 rotation IV

D.1 Disassembly path with the energy minimum protein combination obtained for each PDB structure of the Adeno-Associated Virus, Avian Birnavirus, Bovine Parvovirus, Human Parvovirus, Porcine Parvovirus and Rodent Protoparvovirus groups, in each step of the removal of one protein, with heuristic II. XXIII
D.2 Disassembly path with the energy minimum protein combination obtained for each PDB structure of the *Bombyx mori* and *Galleria mellonella* Densoviruses groups (A), and *Penaeus stylirostris* Densovirus groups (B), in each step of the removal of one protein, with heuristic II. XXIV
D.3 Disassembly path with the energy minimum protein combination obtained for each PDB structure of the Canine and Feline Panleukopenia Virus group in each step of the removal of one protein, with heuristic II. XXV
D.4 Disassembly path with the energy minimum protein combination obtained for each PDB structure of the Hepatitis E Virus group in each step of the removal of one protein, with heuristic II. XXVI
D.5 Disassembly path with the energy minimum protein combination obtained for each PDB structure of the Human Adenovirus group, in each step of the removal of one protein, with heuristic II. XXVII
D.6 Disassembly path with the energy minimum protein combination obtained for each PDB structure of the Porcine Circovirus group in each step of the removal of one protein, with heuristic II. XXVIII

D.7 Disassembly path with the energy minimum protein combination obtained for each PDB structure of the Satellite Tobacco Mosaic Virus group in each step of the removal of one protein, with heuristic II. . . XXIX

E.1 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Adeno-Associated Virus group in each step of the removal of one protein, with heuristic III. XXXIV

E.2 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the *Bombyx mori* (A), *Galleria mellonella* (B), *Penaeus stylirostris* (C) Densoviruses groups, in each step of the removal of one protein, with heuristic III. XXXV

E.3 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Avian Birnavirus, Bovine Parvovirus, Human Parvovirus, Porcine Parvovirus and Rodent Protoparvovirus groups, in each step of the removal of one protein, with heuristic III. XXXVI

E.4 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Canine and Feline Panleukopenia Virus group in each step of the removal of one protein, with heuristic III.XXXVII

E.5 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Hepatitis E Virus group in each step of the removal of one protein, with heuristic III. XXXVIII

E.6 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Human Adenovirus group in each step of the removal of one protein, with heuristic III. XXXIX

E.7 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Porcine Circovirus group in each step of the removal of one protein, with heuristic III. XL

E.8 Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Satellite Tobacco Mosaic Virus group in each step of the removal of one protein, with heuristic III. . . XLI

List of Tables

2.1 Donor and Acceptor's Atoms used to calculate Hydrogen Bonds. . . . 10
2.2 PDB_{IDs} of $T = 1$ Capsid Structures divided by their groups. Viruses marked with † are Parvoviruses (Family *Parvoviridae*). 17

4.1 Ratio of contacts between proteins $[22 \leftrightarrow 23]$ and $[10 \leftrightarrow 23]$. 44

A.1 Symmetry classes included in some symmetry groups. III
A.2 Cyclical Permutation on the C_4 symmetry group and their Cycle Indexes. VI

C.1 Results when five proteins are removed, for Adeno-Associated Virus Capsid structures, using the heuristic I. XV
C.2 Results when five proteins are removed, for Bovine, Human and Porcine Parvoviruses Capsid structures, using the heuristic I. XVI
C.3 Results when four ($N = 4$) and five ($N = 5$) proteins are removed, for *Bombyx mori*, *Galleria mellonella* and *Penaeus stylirostris* Densoviruses Capsid structures, using the heuristic I. XVI
C.4 Results when five proteins are removed, for Hepatitis E Virus Capsid structures, using the heuristic I. XVI
C.5 Results when five proteins are removed, for Canine and Feline Panleukopenia Virus Capsid structures, using the heuristic I. XVII
C.6 Results when five proteins are removed, for Rodent Protoparvovirus Capsid structures, using the heuristic I. XVII
C.7 Results when five proteins are removed, for Satellite Tobacco Mosaic Virus Capsid structures, using the heuristic I. XVII
C.8 Results when five proteins are removed, for Avian Birnavirus Capsid structures, using the heuristic I. XVIII

C.9 Results when five proteins are removed, for Porcine Circovirus Capsid structures, using the heuristic I. XVIII
C.10 Results when five proteins are removed, for Human Adenovirus Capsid structures, using the heuristic I. XVIII

D.1 Results when five proteins are removed, for Adeno-Associated Virus Capsid structures, using the heuristic II. XIX
D.2 Results when five proteins are removed, for Bovine, Human and Porcine Parvoviruses Capsid structures, using the heuristic II. XX
D.3 Results when four ($N = 4$) and five ($N = 5$) proteins are removed, for *Bombyx mori*, *Galleria mellonella* and *Penaeus stylirostris* Densoviruses Capsid structures, using the heuristic II. XX
D.4 Results when five proteins are removed, for Hepatitis E Virus Capsid structures, using the heuristic II. XX
D.5 Results when five proteins are removed, for Canine and Feline Panleukopenia Virus Capsid structures, using the heuristic II. XXI
D.6 Results when five proteins are removed, for Rodent Protoparvovirus Capsid structures, using the heuristic II. XXI
D.7 Results when five proteins are removed, for Satellite Tobacco Mosaic Virus Capsid structures, using the heuristic II. XXI
D.8 Results when five proteins are removed, for Avian Birnavirus Capsid structures, using the heuristic II. XXI
D.9 Results when five proteins are removed, for Porcine Circovirus Capsid structures, using the heuristic II. XXII
D.10 Results when five proteins are removed, for Human Adenovirus Capsid structures, using the heuristic II. XXII

E.1 Results when five proteins are removed, for Adeno-Associated Virus Capsid structures, using the heuristic III. XXX
E.2 Results when five proteins are removed, for Bovine, Human and Porcine Parvoviruses Capsid structures, using the heuristic III. XXXI
E.3 Results when four ($N = 4$) and five ($N = 5$) proteins are removed, for *Bombyx mori*, *Galleria mellonella* and *Penaeus stylirostris* Densoviruses Capsid structures, using the heuristic III. XXXI

E.4 Results when five proteins are removed, for Hepatitis E Virus Capsid structures, using the heuristic III. XXXI
E.5 Results when five proteins are removed, for Canine and Feline Panleukopenia Virus Capsid structures, using the heuristic III. XXXII
E.6 Results when five proteins are removed, for Rodent Protoparvovirus Capsid structures, using the heuristic III. XXXII
E.7 Results when five proteins are removed, for Satellite Tobacco Mosaic Virus Capsid structures, using the heuristic III. XXXII
E.8 Results when five proteins are removed, for Avian Birnavirus Capsid structures, using the heuristic III. XXXIII
E.9 Results when five proteins are removed, for Porcine Circovirus Capsid structures, using the heuristic III. XXXIII
E.10 Results when five proteins are removed, for Human Adenovirus Capsid structures, using the heuristic III. XXXIII

Chapter 1

Introduction

1.1 Virus

Viruses are nucleoproteic intracellular parasites that replicate inside a living cell, infecting it and using its genetic and protein synthesis machinery to create new copies (1; 2; 3; 4). Viruses can infect a wide variety of hosts, from eukaryotic to prokaryotic cells, from singular to multicellular life forms (2).

Viral particles are composed by a nucleic acid (DNA or RNA, single or double stranded) and a capsid, a proteic structure to protect the genetic information in between infections (2; 5; 6). Sometimes viruses are enveloped by a lipid bilayer (3; 6) carrying proteins and/or glicoproteins of previous hosts (7), easing the attachment to the host cells and therefore their infection.

Host cells have many barriers against the infection of viruses, specially the physical barriers, such as cell walls and membranes (2; 4). The method of entry differs not only on the virus species but also on the host cell type, as most entries depend on the attachment and signaling through membrane receptors of the host cell (4; 7; 8; 9). Frequently, the penetration of the cell membrane is done through endocytosis (4; 8). From the endocytosis vesicle the virus can escape or be directed to an endosome, where the natural pH change of this compartment can induce the penetration of the membrane (4). Enveloped viruses have a simpler way of penetrating the membrane, since they only need to fuse their outter membrane with the host's membrane, delivering the viral content directly to the cytosol (2; 4). To deliver the genetic information and thus infect, the virus needs to uncoat the nucleic acid from the capsid, so it can be used by the cell's

machinery to produce other viruses. This uncoating can be done in the cytosol or inside the nucleus (2; 3; 4; 10)

Humans can be infected by a wide range of viruses, leading to multiple deaths all around the world. Influenza Virus, commonly know as the flu, leads to the death of 250 million to 500 million people every year (11). Hepatitis Viruses cause 1.4 million deaths every year, of which 47% are attributed to Hepatitis B Virus (HBV) infections and 48% to Hepatitis C Virus (HCV) infections (11). Human Immunodeficiency Virus (HIV) affects 0.8% of the world population, between the ages of 15 to 49 (11). In 2015, there were 36.7 million people infected with HIV and there were 1.1 million deaths (11). Viruses' related pathologies are, clearly, a major human health problem. This also makes viruses a scientific object of study. Virus research encompasses not only the different ways of infection and the pathologies induced, but also their structure, in order to find a way to fight human viral infections. This work will focus on the disassembly process of virus capsids, with the goal of providing insights on this process that can be used in the development of antiviral drugs that might tackle intermediate states of the process. Moreover, these insights might contribute to the rational design of artificial virus-like nano-particles of biomedical interest (12).

1.2 Capsid Structure

Viruses' capsids can be classified in three main groups according to the overall shape: helical, round and complex (5; 6; 13). On helical structures, the proteins of the capsid wind around the genetic material, making a rod-shaped virus in which the length varies depending on the size of the genetic material (13). An example of this type of viruses is the Tobacco Mosaic Virus (TMV), which its structure is represented on Figure 1.1.

On the other hand, round-shaped viruses have a fixed number of proteins or asymmetric units (which can be composed of one or more proteins) surrounding the genetic material in a icosahedral symmetry (5; 6). The arrangement of theses proteins was first described by Caspar and Klug in 1962 (5), who developed a method to classify the icosahedral symmetry by triangulating the icosahedron facets. The Triangulation number (or T-number) represents the number of equilateral triangles in which a triangular face of the icosahedron is divided into (6).

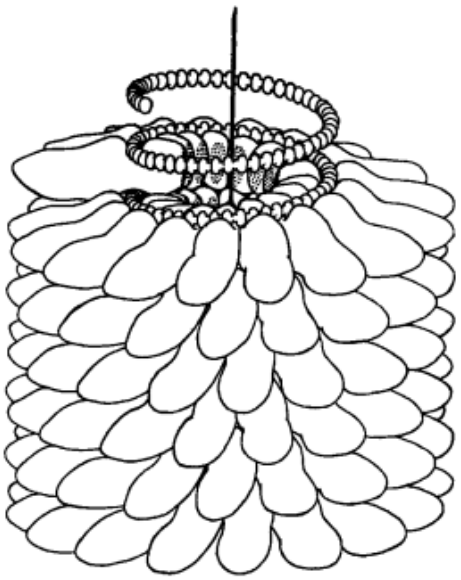

Figure 1.1: Tobacco Mosaic Virus structure, with the nucleic acid in the middle surrounded by the capsid proteins, as drawn by Klug and Caspar in 1960 (14).

The T-number is equal to $T = Pf^2$, in which $P = h^2 + hk + k^2$, where h and k are integers with no common factors, resulting in the series $[1,3,7,13,19,21,31,37,\ldots]$, and $f \in \mathbb{N}$. (5; 6). For the same value of f, we can have many results of T, depending on the value of P, resulting in different classes. For $P \geq 7$, the structure starts to be skewed, having two different symmetries: a right-handed (*dextro*) and a left-handed (*laevo*) symmetry.

The number of proteins in T-numbered virus is $S = 60 \times T$ and the number of pentamers and hexamers formed by the arrangement of the proteins on the capsid $M = 10(T-1)$ *hexamers* $+ 12$ *pentamers*. (5; 6) On this thesis, we will focus on $T = 1$ icosahedral capsid structures.

1.3 Virus Assembly and Disassembly Studies

Many studies have addressed the virus assembly process, both experimentally and theoretically (reviewed in (13) and (15)). Viruses' capsid subunits are held together mainly by weak contacts (16; 17; 18) such as electrostatic, van der Waals and hydrogen bonds. The consequence of these weak interactions is a globally stable capsid (19).

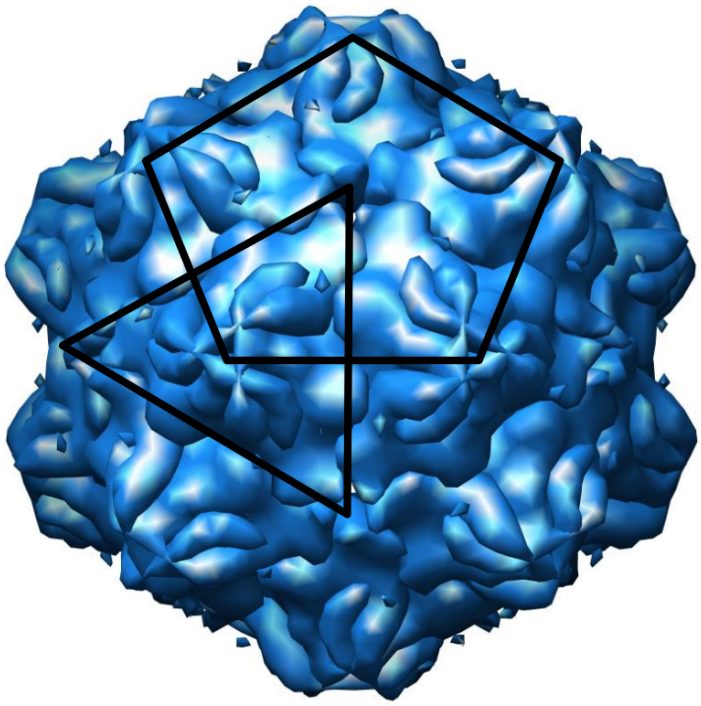

Figure 1.2: Canine Panleukopenia Virus Capsid. $T = 1$ structure, where the triangle represents the face of a icosahedron, and the pentagon represents the pentamers around a 5-fold axis. Structure PDB_{ID} $1c8h$.

The assembly is a spontaneous process *in vitro* under close-to-physiological conditions (15) and the resulting capsids are sufficiently stable for a long period of time.

Assembly has sigmoidal kinetics (19), starting with a lag phase where intermediates with different degrees of complexity, ranging from the individual subunits to the complete capsid, are formed, until they reach a steady-state concentration. The concentration of intermediates is rather low, transient and short-lived, since its accumulation would lead to high concentrations of intermediates states, depleting all reserves of subunits needed to reach the final capsids (19). The dissociation of these intermediates is the rate-limiting step of this kinetic (19), being essential a higher rate of dissociation on the first steps than of association.

This leads to a nucleation, where there is an intermediate which is both prone to disassemble completely into its individual subunits, or to assembly into the complete capsid (see Figure 1.3). This limits the possibilities of divergent assembly pathways that can consume all the individual subunits, resulting in a more robust assembly (19).

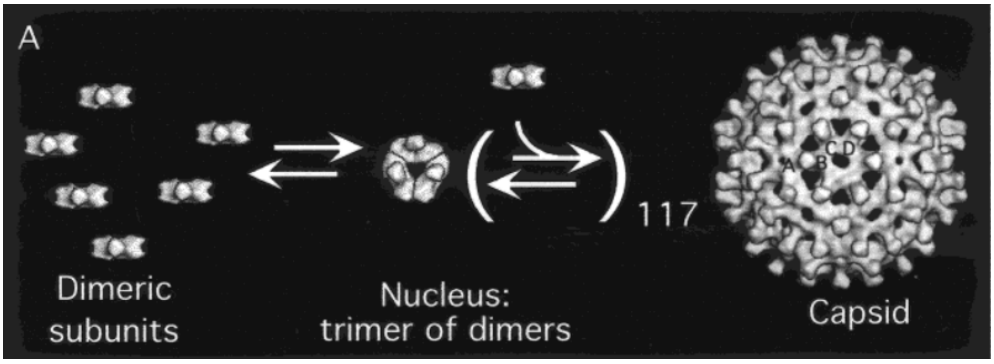

Figure 1.3: Scheme of the Hepatitis B Virus capsid ($T = 4$) assembly, with the formation of a nucleus, followed by the elongation by adding multiple dimeric subunits to form a complete capsid. Adapted from (17).

After reaching a steady-state of intermediates, the formation of the complete capsid enters an exponential phase. On this stage there is an addition of subunits or clusters of subunits to a small nucleus of assembly. The equilibrium is reached when only complete capsid and free subunits are found (19).

Many theoretical models were built to describe the virus assembly, which can be applicable also to disassembly. Independently of the approach, Zlotnick (16; 17), Reddy (20; 21) and Rapaport (22; 23) derived similar conclusions for different viruses' capsids.

Reddy et al. (20; 21), through a theoretical approach, reveals a low energy minimum before the complete capsid is assembled, without a pentamer of triangles. This also follows the idea suggested that a trimeric protomer reservoir is readily formed, so that the assembly steps are done by the addition of such trimers (20) . The studies of Rapaport et al. (22; 23) show that long-lived transient structures of capsids with just one last triangle of proteins missing to form the complete capsid are very abundant, agreeing with the results of Reddy et al.

The combinatorial method used to evaluate if subunit complexes were identical by Reddy et al. (20) and Horton et al. (24) was based on the calculation of all the possible combinations and evaluation, one by one, of those which are identical. However, these authors never actually specify which are the criteria behind their evaluation, suggesting the usage of symmetry operations to the structure, but never going into detail. In two studies, the stability of the different intermediary structures is evaluated (20; 24), using

molecular dynamics or a heuristical measurement based on the variation of exposed surface to the solvent, before and after the formation of a complex.

In high contrast to the number of studies on the assembly process, there are very few studies with experimental results of the disassembly pathway. Since it is expected that intermediates of this process are transient and very difficult to detect experimentally on such a short period, the disassembly pathways are always difficult to be confirmed experimentally. Moreover, assembly and disassembly pathways might be distinct since some viruses undergo a maturation step after complete assembly, such as proteolysis, crosslinking or conformational change (19).

Castellanos et al. (12), using Atomic Force Microscopy (AFM), observed the removal of a triangle of subunits, formed by three proteins (an icosahedral face), when force was applied over a Minute Virus of Mice (MVM) capsid, followed sometimes by the removal of an adjacent triangle. In other cases, a full removal of a pentamer of triangles was observed, with 15 proteins of the capsid being removed. All this follows the prediction of Reddy et al. (20; 21), Rapaport et al. (22; 23) and Horton et al. (24) which the latter states that the energy minimums appear every multiple of three subunits.

This work will tackle the lack of theoretical studies directed specifically on the disassembly pathway and with a real combinatorial background, with all mathematical insight behind, to explain it into detail.

1.4 Aims

The aim of this study was to understand the disassembly path taken by virus capsids, through a computational approach. This work uses a novel approach to predict the process of disassembly of a virus capsid, integrating many ideas of geometry and combinatorics, since most studies in structural virology do not take these into account, or do not develop deeply enough into polyedric combinatorics (such as (20; 24)). Another unique feature of this approach is the fact that, unlike Zlotnick's combinatorial factors (16), used in the prediction of assembly pathways which are based on a Dodecahedron, and not on the full capsid, the computational model developed for this dissertation is the Deltoidal Hexecontahedron, a 60-faced polyhedron.

The working hypothesis supporting this dissertation is:

Hypothesis: There is a common process of capsid disassembly among the different families of virus with icosahedral symmetry and, if not, we can identify the virus family based on its particular disassembly path.

Another goal of this study is to propose and compare different heuristical measures of the total energy of a given capsid fragment. This approach that conjugates geometrical aspects with carefully selected energy metrics supports the long term objective of predicting the disassembly path by a computational approach, being possible to predict targets for antiviral drugs that can interfere with the disassembly process.

Chapter 2

Methods

2.1 Combinatorics and Symmetry

Combinatorially, the problem of removing protein subunits from an icosahedral viral capsid is equivalent to painting the faces of an icosahedron with two colours, representing 'present' or 'absent'. Therefore, the number of different ways to remove a given number N of subunits, is the number of different combinations obtained by choosing N faces to be painted with one colour out of a total of 20 faces of the icosahedron. These can be computed using Equation A.1. However, although the virus capsid structures with $T = 1$ have icosahedral symmetry, they are not icosahedron-like, as they contain 60 proteins instead of only 20. The 60-face polyhedron, in which each triangular face of the icosahedron is divided into three faces, is the Deltoidal Hexecontahedron (Figure 2.1).

Due to the symmetry of the Deltoidal Hexecontahedron, some of the combinations generated would be symmetrically equivalent, and the real number of combinations is, actually, less than the predicted $2^{60}\left(=\sum_{n=0}^{60}\binom{60}{n}\right)$. The exact number of combinations can be obtained by studying the Burnside's Lemma (26). [1] A symmetry group is a mathematical group where its elements are symmetry operations on an object. A symmetry group for the Deltoidal Hexecontahedron was built on this work to study the different symmetries of this polyhedron. Furthermore, the proteins that make up a triangular facet on an icosahedral capsid structure (Figure 1.2) can be arranged in the

[1] This lemma, and other mathematical concepts supporting this thesis, are presented in Appendix A.

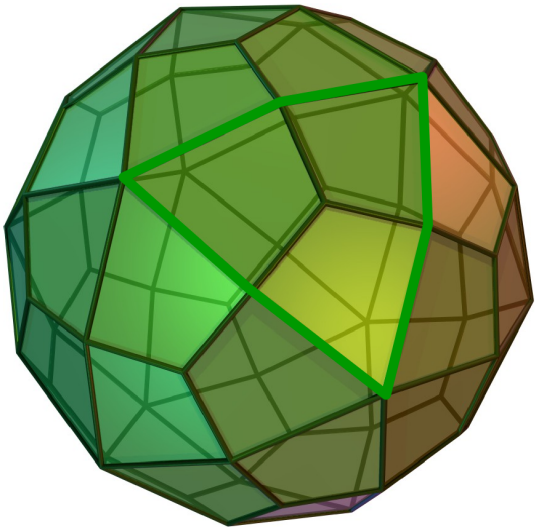

Figure 2.1: Deltoidal Hexecontahedron, from (25). The highlighted region represents a triangular face of an Icosahedron.

two ways represented in Figure 2.2. However, regardless of their arrangement in each face, the symmetry group described in Appendix B still applies, since the symmetry operations continue to be the same.

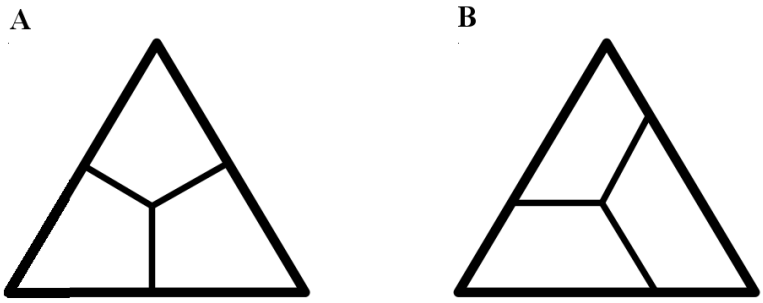

Figure 2.2: Possible arrangements of the proteins on the triangular facets of the viruses capsids in a deltoidal (A) and a trapezoid (B) fashion.

2.2 Structures of Viruses' Capsids and Energy Calculation

All the capsids' structures were obtained from the Protein Data Bank (*PDB*, `http://www.rcsb.org/`, (27)). This work focuses on the analysis of structures of less complexity, namely, $T = 1$ capsids which have only 60 proteins. As such, the database ViperDB (`http://viperdb.scripps.edu`, (28)) was used in order to restrict the collection to the structures with $T = 1$. With the list obtained from the Web API of ViperDB, some Viruses' Capsids with $T = 1$ were rejected due to the fact that they were not divided into a 60-model *PDB* file and due to not using the subunit numbering followed by the majority of the *PDB* files (as seen in Figure 2.3). Structures with missing residue side chains were also excluded. Non-proteic compounds found in the structures were ignored.

To calculate the energy of the full capsid and its disassembly products, the number of hydrogen bonds, salt bridges and hydrophobic contacts between different proteins of the capsid were counted (intra-protein contacts were ignored) .

2.2.1 Hydrogen Bonds

The hydrogen bonds were counted if the acceptor and donor's atoms were at a distance less than or equal to 4.0Å. The atoms (and correspondent amino acids) of each category are indicated in Table 2.1.

Table 2.1: Donor and Acceptor's Atoms used to calculate Hydrogen Bonds.

	Amino acids
Donor	ARG($N\varepsilon$; $N\eta_1$; $N\eta_2$), ASN($N\delta_2$), CYS($S\gamma$), GLN($N\varepsilon_2$), HIS($N\delta_1$; $N\varepsilon_2$), LYS($N\zeta$), SER($O\gamma$), THR($O\gamma_1$), TRP($N\varepsilon_1$), TYR($O\eta$)
Acceptor	ASN($O\delta_1$), ASP($O\delta_1$; $O\delta_2$), GLN($O\varepsilon_1$), GLU($O\varepsilon_1$; $O\varepsilon_2$), HIS($N\delta_1$; $N\varepsilon_2$), SER($O\gamma$), THR($O\gamma_1$), TYR($O\eta$)

2.2.2 Salt Bridges

For every acidic amino acid (ASP and GLU), the existence of any basic amino acid (ARG and LYS) in the radius of 14.0Å of the β-carbon was analysed, and, if so, it was counted the existence of a bond only if their positive charged atoms ($O\varepsilon_i$ or $O\delta_i$) were in the radius of 4.0Å of the negative charged atoms ($N\zeta$ or $N\eta_i$).

2.2.3 Hydrophobic Contacts

The number of hydrophobic contacts between capsid proteins was calculated as the number of times that the distance between β-carbons was less than or equal to 7.0Å for every two residues of the following amino acids: ALA, VAL, LEU, ILE, MET, PHE, TYR and TRP.

2.2.4 Energy Calculation

Based on the number of contacts, we can calculate a heuristic measure of total energy for a complete capsid and each of its putative disassembled products. The different measures that were used are described in Equations (2.1) to (2.3). For the sake of simplicity, Equations (2.1) to (2.3) will be denoted further on as heuristic I, II and III, respectively.

$$N_{SB} = Number\ of\ Salt\ Bridges$$
$$N_{HB} = Number\ of\ Hydrogen\ Bonds$$
$$N_{HC} = Number\ of\ Hydrophobic\ Contacts$$

$$E_1 = 20 \times N_{SB} + N_{HB} + N_{HC} \tag{2.1}$$

$$E_2 = N_{SB} + N_{HB} + N_{HC} \tag{2.2}$$

$$E_3 = 100 \times N_{SB} + 10 \times N_{HB} + N_{HC} \tag{2.3}$$

Heuristic I gives Salt Bridges 20 times more energy than the Hydrogen Bonds, since an ion-ion interaction has an approximated strength of $250kJ/mol$ (29) while a Hydrogen Bond has an approximated strength of $10kJ/mol$ (29). The gain from making Hydrophobic Contacts was considered to be energetically equivalent to making Hydrogen Bonds, since when hiding hydrophobic amino acids, a water "cage" is destroyed, breaking Hydrogen Bonds made by the water itself to hold this cage (29). In heuristic II, only the total number of inter-subunit bonds in the capsid complex is taken into account; the three types of bounds are equally weighted. On the other hand, on heuristic III an increasing power of 10 was given to each type of inter-subunit bond, in the known order of strength of this types of contacts (Salt Bridges > Hydrogen Bonds > Hydrophobic Contacts) (29).

2.3 Symmetry Group Construction

To build up the group of permutations associated with the rotations of a Deltoidal Hexecontahedron, the faces were numbered as shown in Figure 2.3, followed by the analysis of the transformations done on the numbering of the faces by the symmetry operators of the symmetry group I (Equation 2.4 from Table A.1).

$$I = \left\{E,\ 12C_5,\ 12C_5^2,\ 20C_3,\ 15C_2\right\} \tag{2.4}$$

Applying the different rotations on the symmetry axes of a Deltoidal Hexecontahedron (five, three and two-fold axes, as seen in Figure 2.3), results in a permutation group, shown in Appendix B. This group has 60 permutations, as in the symmetry group I, and is isomorphic to the Alternating Group A_5. These permutations can be seen as permutations of faces of the Deltoidal Hexecontahedron, or vertices of its dual, the Rhombicosidodecahedron. In fact, if the faces of the Deltoidal Hexecontahedron are represented as a graph, with the faces as the vertices and adjacent faces connected by edges, we get the same graph that is obtained by representing the geometrical vertices of the Rhombicosidodecahedron as graph vertices, with the geometrical edges connecting the same vertices on the graph (Figure 2.4). Thus, we can always use either perspectives. Since the graph of a Rhombicosidodecahedron allows the study of each

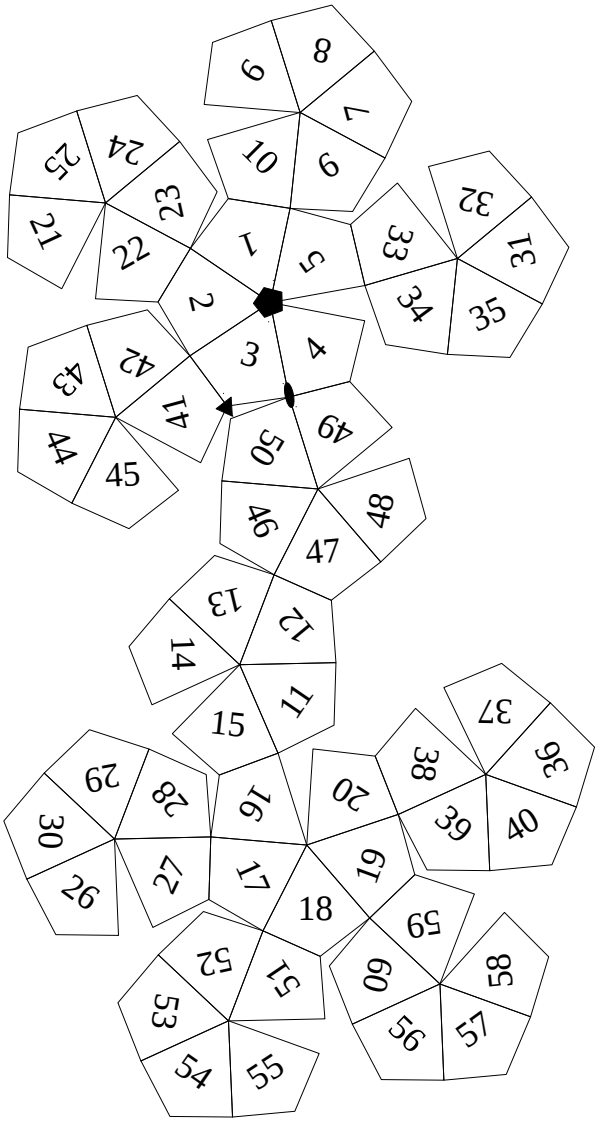

Figure 2.3: Numbered faces of the Deltoidal Hexecontahedron as the proteins of a $T = 1$ virus capsid as found on Protein Data Bank. Five, three and two-fold axes marked with a pentagon, triangle and ellipse, respectively.

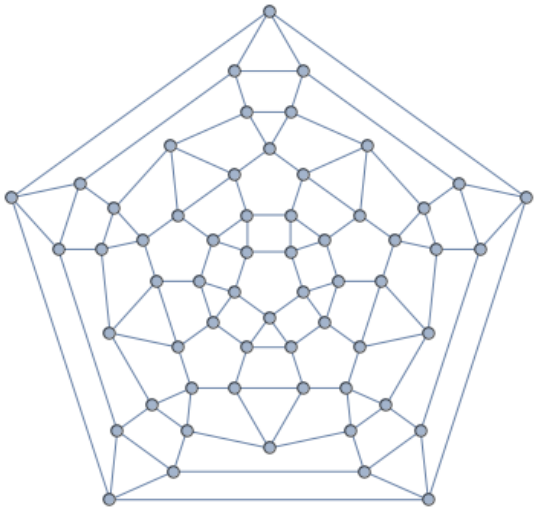

Figure 2.4: Geometrical vertices of the Rhombicosidodecahedron represented as graph vertices, with edges connecting the vertices that are adjacent by geometrical edges. This graph is equivalent to the graph of the faces of the Deltoidal Hexecontahedron represented as vertices and connected with the four adjacent faces by edges.

protein as a vertex, in what follows we mainly used that perspective to develop the following methods of the analysis of virus disassembly.

2.4 Removal of Proteins and Graph Representation

Each possible fragment can be represented by a 60 element-vector of ones and zeros, depending on the presence or absence of each protein with the same number as the index of the position of the ones and zeros. To compute the structures of all the fragments obtained by the removal of N proteins, we first computed the possible combinations of having an N number of zeros, fixing the first position always as zero, to reduce runtime by lowering the combinations to just $\binom{59}{N}$. A list of non-redundant combinations was then constructed as follows: for every combination created, the permutations from the permutation group of Appendix B were applied onto it and if any new combination generated by a permutation was identical to one already present in the list, that first combination would be ignored. If no permutation generated a combination already on the list, the first would be saved on that same list. In the end, non-redundant combinations were obtained.

As said previously, the graph representation was the graph of the vertices of the Rhombicosidodecahedron, where each vertex was associated to a protein of the capsid. The graph representation allowed us to study particular situations exclusive to the problem of removing proteins of capsids, since it shows that every vertex is connected to four other vertices and it depends on their presence to stay on the structure (as seen on Figure 2.4). It is important to remember that the Burnside's Lemma only takes into account the painting of faces, not taking into consideration the fact that, if a protein is not surrounded by the four proteins that should hold it (i.e. surrounded by proteins that are painted as 'absent'), the first will be free to exit the capsid, and it'll be considered as another vertex that is removed from the graph.

Under this principle, all the combinations obtained by the previous step were represented as a graph and, in case there was a vertex or a group of vertices disconnected from the major cluster, they were removed as well and a new combination was created, and it was repeated the process of verifying if it was not redundant. For each combination, the energy was calculated based on the set of proteins present.

2.5 Optimal Path for Disassembly

For all the possibilities of removing a N number of proteins, a tree of all the paths that the virus disassembly process could take was built, from the intact virus capsid to a capsid with $60-N$ proteins, going through every intermediate state. For a better understanding, the combination of proteins that were removed from the structure was analysed, but not those still present. Two vertices of the path of the tree were connected only between combinations where the smallest of the two was a subset of the other, and their sizes' difference was equal to one (e.g. connecting the combination where proteins $[1,2]$ were removed to the combination where proteins $[1,2,3]$ were removed). These edges were given a transition energy weight, based on the energies of each complex and their size, by the Equation 2.5. The size of complex i is equal to $60-N_i$, where N_i in the number of proteins removed on complex i.

$$\Delta E = E_{transition\ i \to j} = \frac{E_j}{size_j} - \frac{E_i}{size_i} \tag{2.5}$$

The Bellman-Ford algorithm was used on this tree to calculate the shortest path from the complete viral capsid to every possibility of a capsid with $60-N$ proteins, based on the weights of the edges of the tree, and recorded the five shortest paths for every N and the corresponding structure(s) responsible for them.

2.6 Graph Representation

The capsid structures that were considered in this work were divided into groups according to the criteria of similarity of the infection host and the genes codifying the proteins of the capsid. The list of results of such grouping is presented in Table 2.2, with 51 structures divided into 14 groups.

For each structure, it was studied the minimum path to arrive to any structure with $60-N$ proteins (with N ranging from 0 to 5) and it was recorded the one with the shortest path (energy minimum), giving it a weight of one. In case there were more than one with the same energy minimum, all would be saved and given a $1/k$ weight to each, where k is the number of combinations with the same minimum. The other four shortest paths (second to fifth energy minimums) were saved as well, but not represented as graphs.

The results obtained were represented as a graph, where the path for each group was drawn, and two vertices were connected when they were consecutive in a path for each *PDB* structure. To each vertex was given a size proportional to the sum of all the weights in different structures of the same group that crossed the same vertex, following the proportion described by the formula in Equation 2.6.

$$Size\ of\ vertex = \frac{100 \times weight_{vertex}}{Size_{group}} \tag{2.6}$$

2.7 Implementation

All the methodology previously described and respective results were processed and analysed with programs written in Python 2.7 programming language using scientific computing modules included in the Anaconda distribution (30) (mainly, igraph (31)). All the code used in this work can be seen and downloaded on the github repository: `https://github.com/CAPiedade/Virus-Disassembly`. UCSF Chimera (32)

Table 2.2: PDB_{IDs} of $T = 1$ Capsid Structures divided by their groups. Viruses marked with † are Parvoviruses (Family *Parvoviridae*).

PDB_{ID}	Group	PDB_{ID}	Group
1x9t	Human Adenovirus Pt-Dd	1p5w	Canine and Feline Panleukopenia Parvovirus†
4aqq		1p5y	
4ar2		1c8h	
1wcd	Avian Birnavirus	1c8d	
3jci	Porcine Circovirus	1ijs	
3r0r		2cas	
2ztn	Hepatitis E Virus	4dpv	
2zzq		1c8f	
3hag		1c8g	
1lp3	Adeno-Associated Virus†	1c8e	
3j4p		1fpv	
5egc		4gbt	Rodent Protoparvovirus†
2g8g		4g0r	
3ntt		2xgk	
		1z14	
3ra4		1z1c	
3ra9		1mvm	
3ra8		1s58	Human Parvovirus†
3raa		1k3v	Porcine Parvovirus†
4rso		1a34	Satellite Tobacco Mosaic Virus
2qa0		2buk	
3ra2		4bcu	
3ux1		4oq8	
3j1q		4qc8	Bovine Parvovirus†
4iov		3n7x	*Penaeus stylirostris* Densovirus†
1dnv	*Galleria mellonella* Densovirus†	3p0s	*Bombyx mori* Densovirus†

was used for the visualization and rendering of figures of Viral Capsids presented in this work.

Chapter 3

Results

3.1 Paths of Disassembly

The paths to disassemble the different capsids found in this study are portrayed in Figures 3.1 to 3.10. These paths were obtained using heuristic I.

It can be observed that Adeno-Associated Virus tend to follow paths of disassembly which can be roughly classified into three main groups. Two characterized by the prevalence of structures where subunits $\{1,2,10,22,23\}$ [1], and $\{1,2,6,10,23\}$ (Figure 3.1) are removed as their minimum and one with two energetically identical structures, $\{1,6,25,26,53\}$ and $\{1,6,13,45,46\}$ (Figure 3.1 and Table C.1). Moreover, the most populated intermediate is the combination $\{1,10,23\}$ which, by analysing the Deltoidal Hexecontahedron, is equivalent to removing a triangular face of the Icosahedron (Figure 2.3).

Bovine, Human and Porcine Parvovirus have a very similar path of disassembly (Figure 3.2), being different only when it comes to the Porcine Parvovirus, when four proteins are removed (Figure 3.2). Even so, both paths end up in the same structure with the removal of subunits $\{1,2,10,22,23\}$.

The Densoviruses analysed in this study (*Bombyx mori*, *Galleria mellonella* and *Penaeus stylirostris*) showed some conservation of paths (Figure 3.3), although it can be pointed out the fact that the first and second energy minimums were equal in all the

[1] $\{\alpha_1,\alpha_2,\alpha_3,\ldots,\alpha_N\}$ represents the removal of a set of proteins, with size N, where protein numbered as α_i was removed.

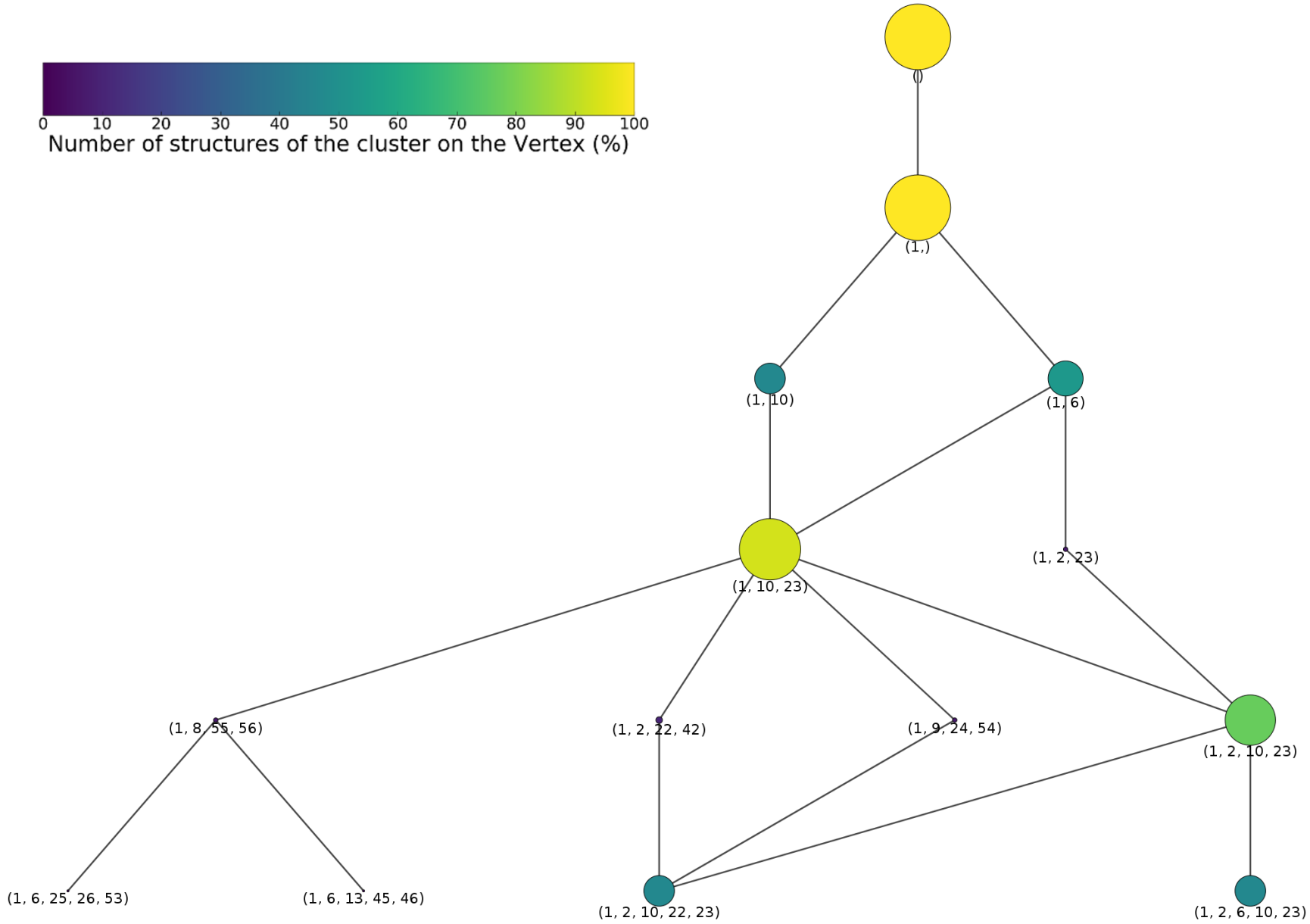

Figure 3.1: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Adeno-Associated Virus group in each step of the removal of one protein, with heuristic I.

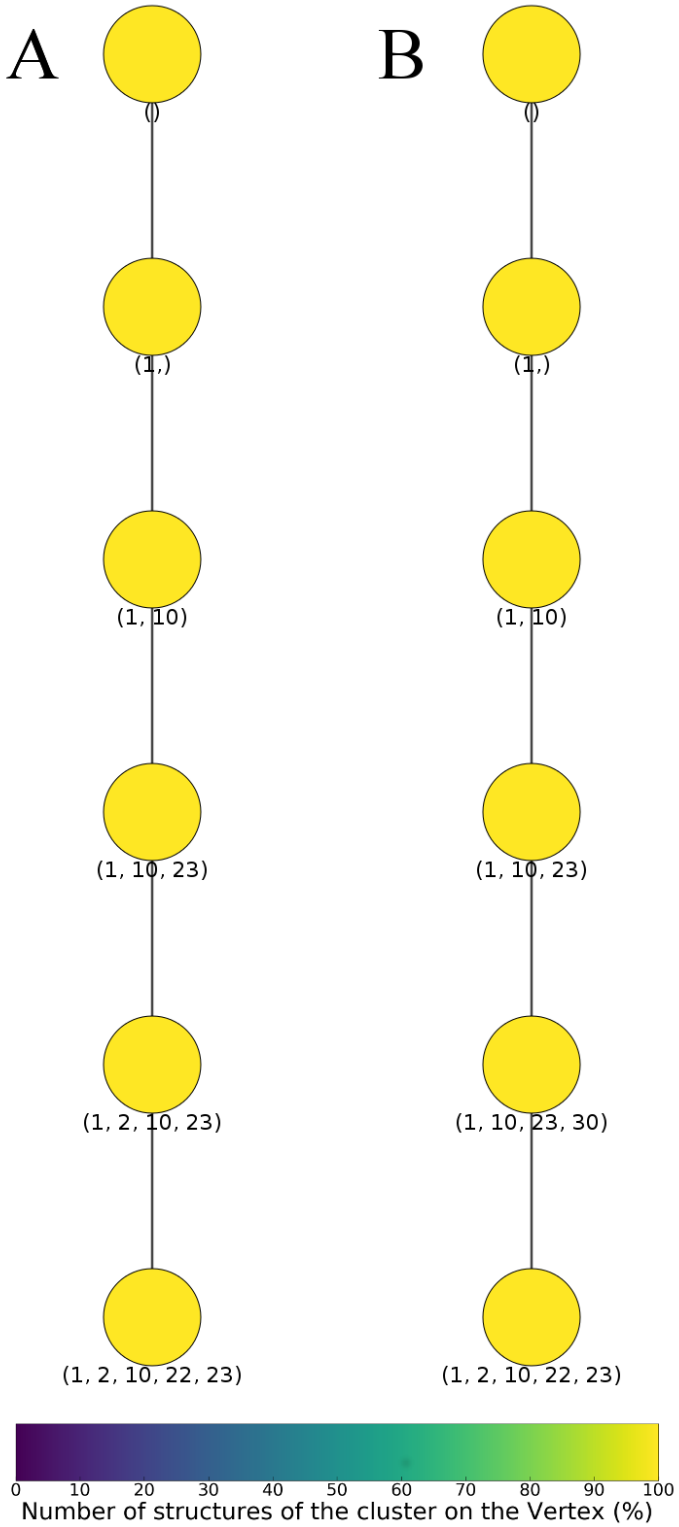

Figure 3.2: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Bovine and Human Parvovirus groups (A), and Porcine Parvovirus (B) group, in each step of the removal of one protein, with heuristic I.

paths, when four proteins were removed: $\{1,2,22,23\}$ and $\{1,2,10,23\}$, respectively (Table C.3).

The Canine and Feline Panleukpopenia Virus, Rodent Protoparvovirus, Porcine Circovirus and Avian Birnavirus (Figures 3.4 to 3.7) followed a main path of disassembly, ending up at the same structure, in which proteins $\{1,2,10,22,23\}$ were removed.

There were two distinct paths followed by the Hepatitis E Virus, one directed to structure $\{1,2,6,10,23\}$, the other to structure $\{1,2,10,22,23\}$ (Figure 3.8). It is interesting to notice that the second energy minimum of structure 3hag (Table C.4) was $\{1,2,6,10,23\}$, which is the first energy minimum of the other two structures, but, conversely, this did not happen for the others.

Human Adenovirus Pt-Dd adopted a path of disassembly around the five-fold axis (Figure 2.3), ending in a structure without the five proteins of the combination $\{1,2,3,4,5\}$ (Figure 3.9).

Finally, the Satellite Tobacco Mosaic Virus took two different paths of disassembly, in which one ended up in a structure without the proteins of the combination $\{1,2,6,10,24\}$, and the other branch ended up splitting into the combinations $\{1,2,3,23,42\}$ and $\{1,2,6,10,23\}$ (Figure 3.10).

3.2 Structural Paths

For all the paths presented on the previous chapter, Figures 3.11 and 3.12 show the proteins removed on different stages of the paths and the resulting Rhombicosadodecahedron graphs.

Analysing the removal of proteins in positions $\{1,10,23\}$ (Figure 3.13) resulted in a triangle-shaped structure, as predicted by the graph on Figure 3.11.B. The removal of proteins $\{1,2,22,23\}$ (Figure 3.14), common among Densovirus structures (3.3), led to a square-shaped entrance to the interior of the capsid (Figure 3.11.C).

According to these results, Parvoviruses groups(Figures 3.1, 3.2, 3.4 and 3.5), Porcine Circovirus group (Figure 3.6) and Avian Birnavirus group (Figure 3.7) usually adopted a capsid where the combination $\{1,2,10,23\}$ was removed, represented in Figure 3.15. This structure is less symmetrical compared to the one of $\{1,2,22,23\}$ (see Figure 3.11.D against Figure 3.11.C)

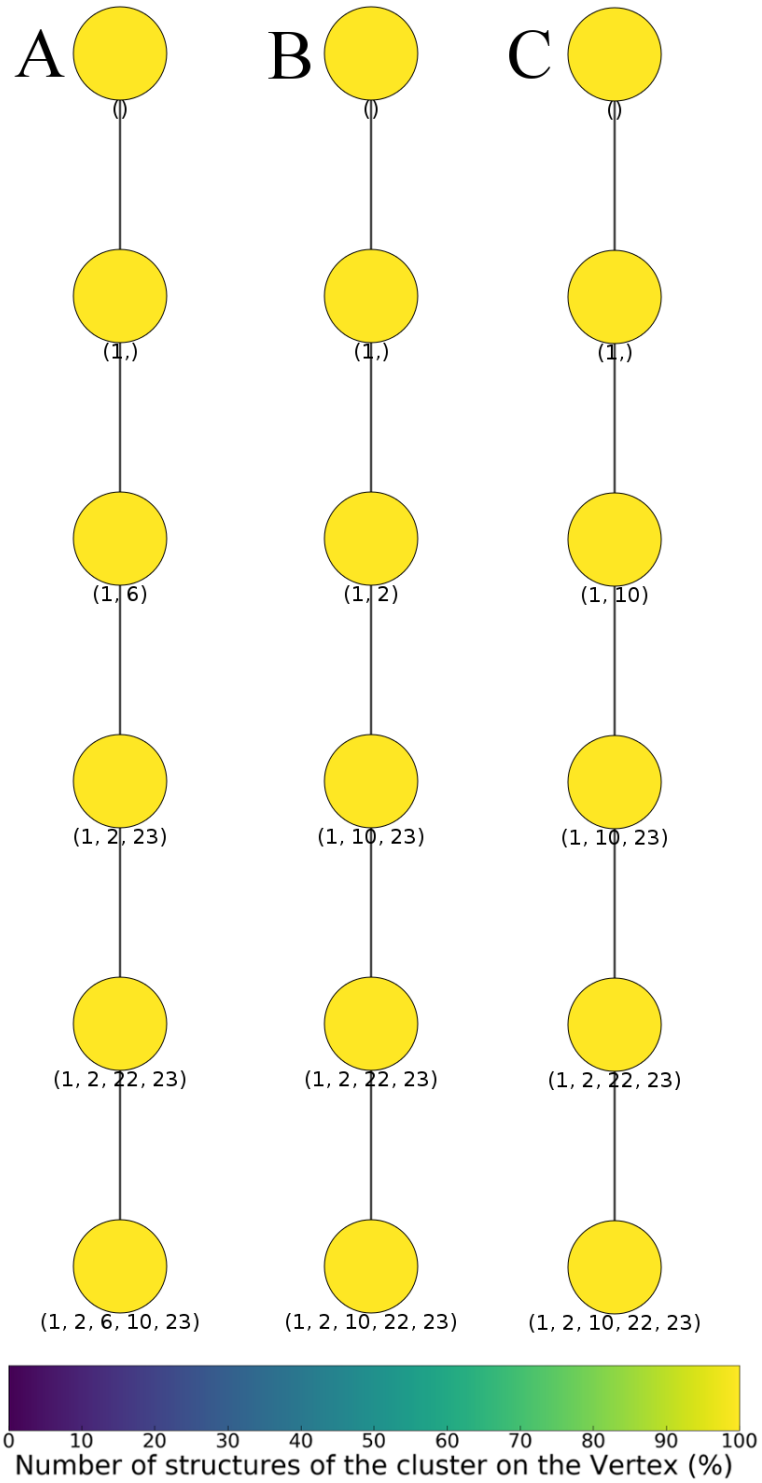

Figure 3.3: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the *Bombyx mori* (A), *Galleria mellonella* (B), *Penaeus stylirostris* (C) Densoviruses groups, in each step of the removal of one protein, with heuristic I.

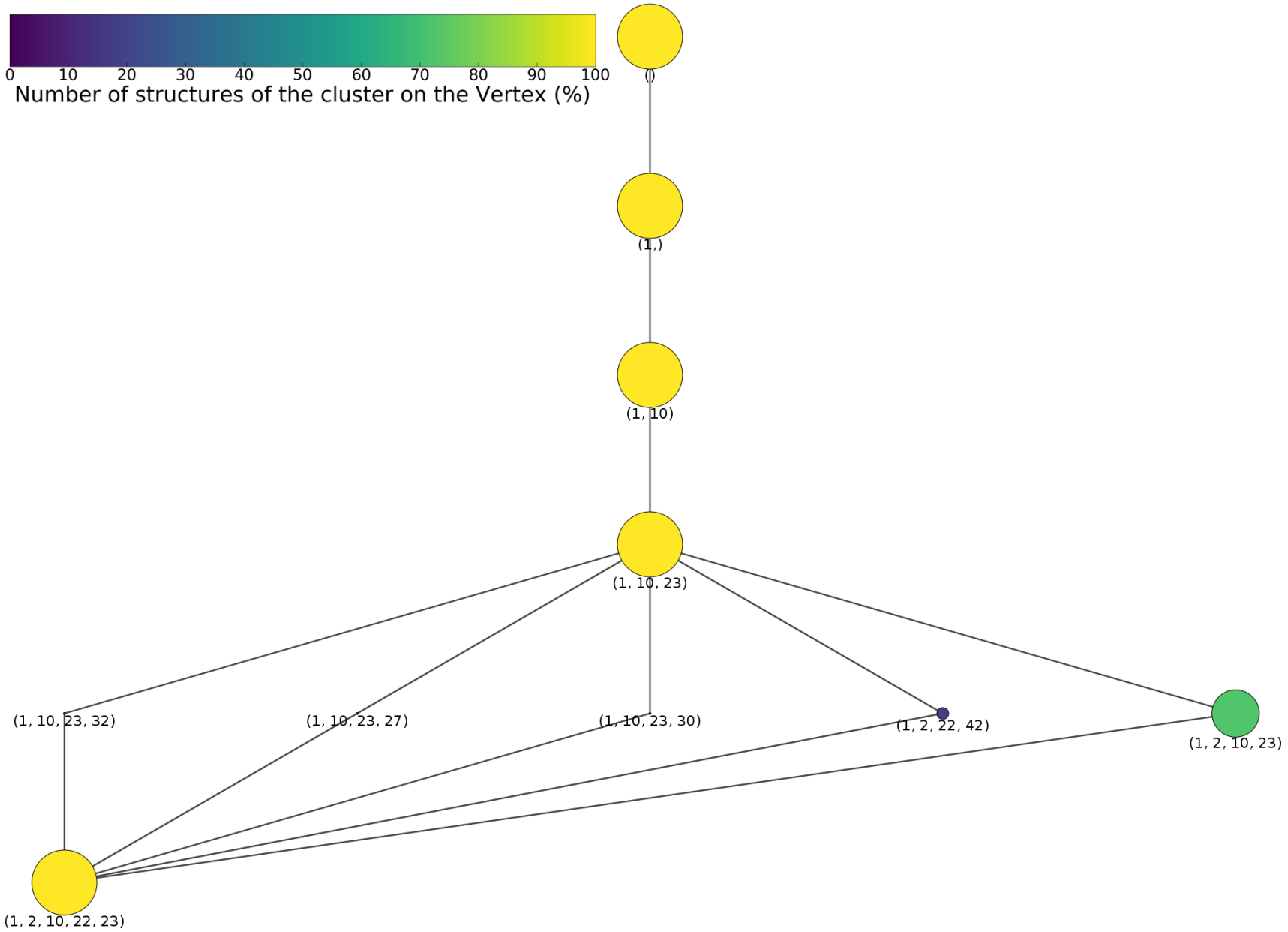

Figure 3.4: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Canine and Feline Panleukopenia Virus group in each step of the removal of one protein, with heuristic I.

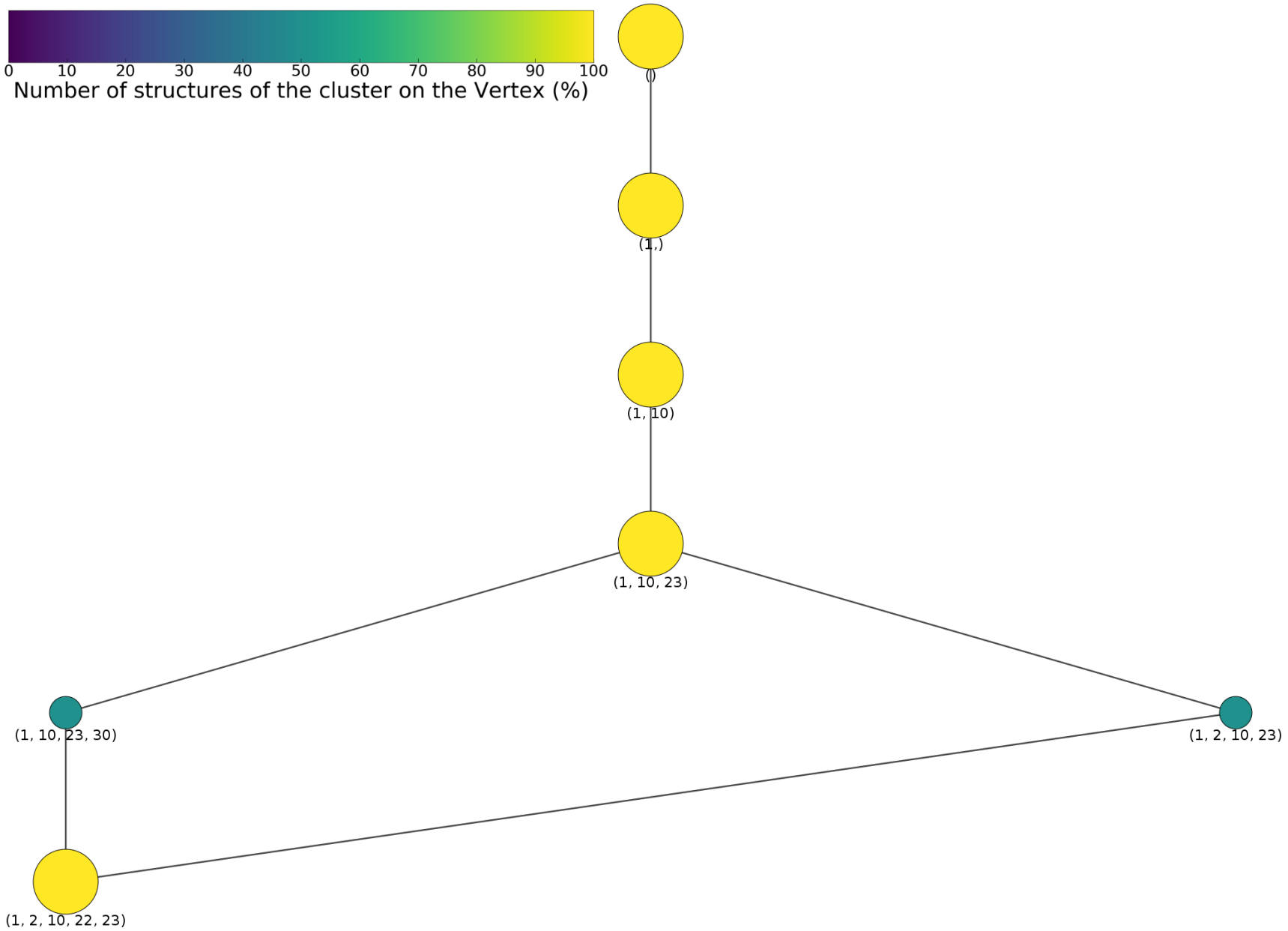

Figure 3.5: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Rodent Protoparvovirus group in each step of the removal of one protein, with heuristic I.

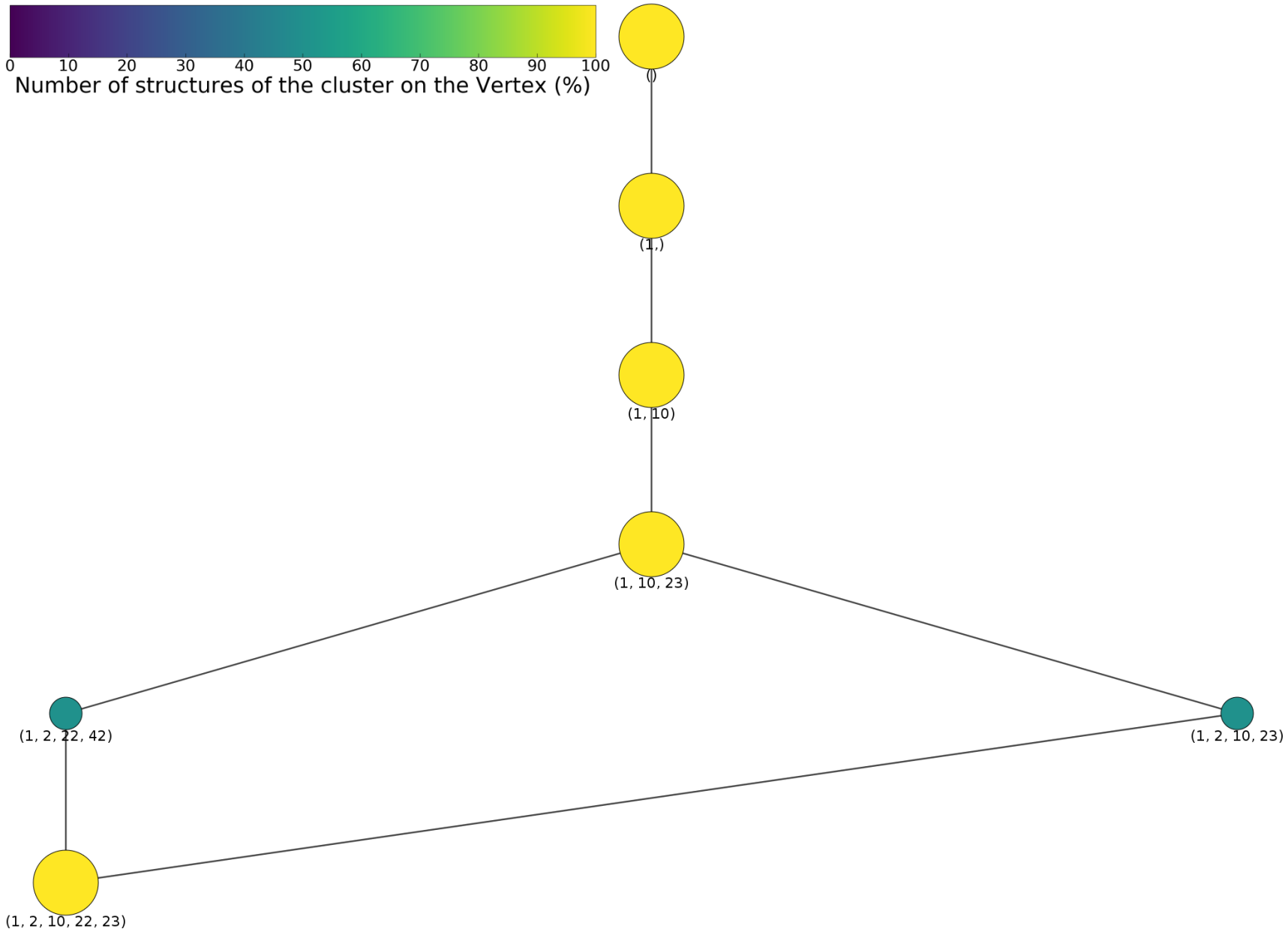

Figure 3.6: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Porcine Circovirus group in each step of the removal of one protein, with heuristic I.

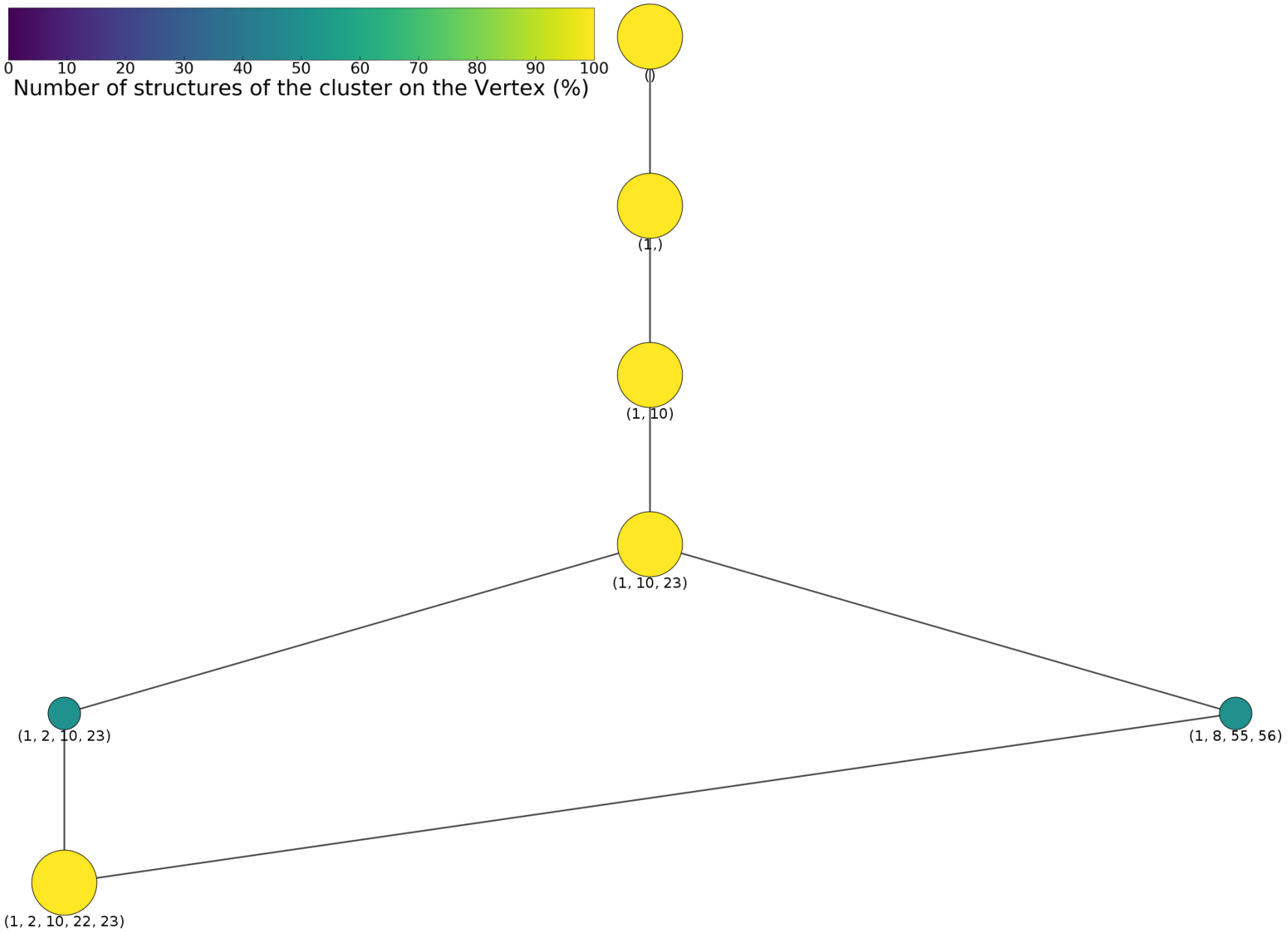

Figure 3.7: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Avian Birnavirus group in each step of the removal of one protein, with heuristic I.

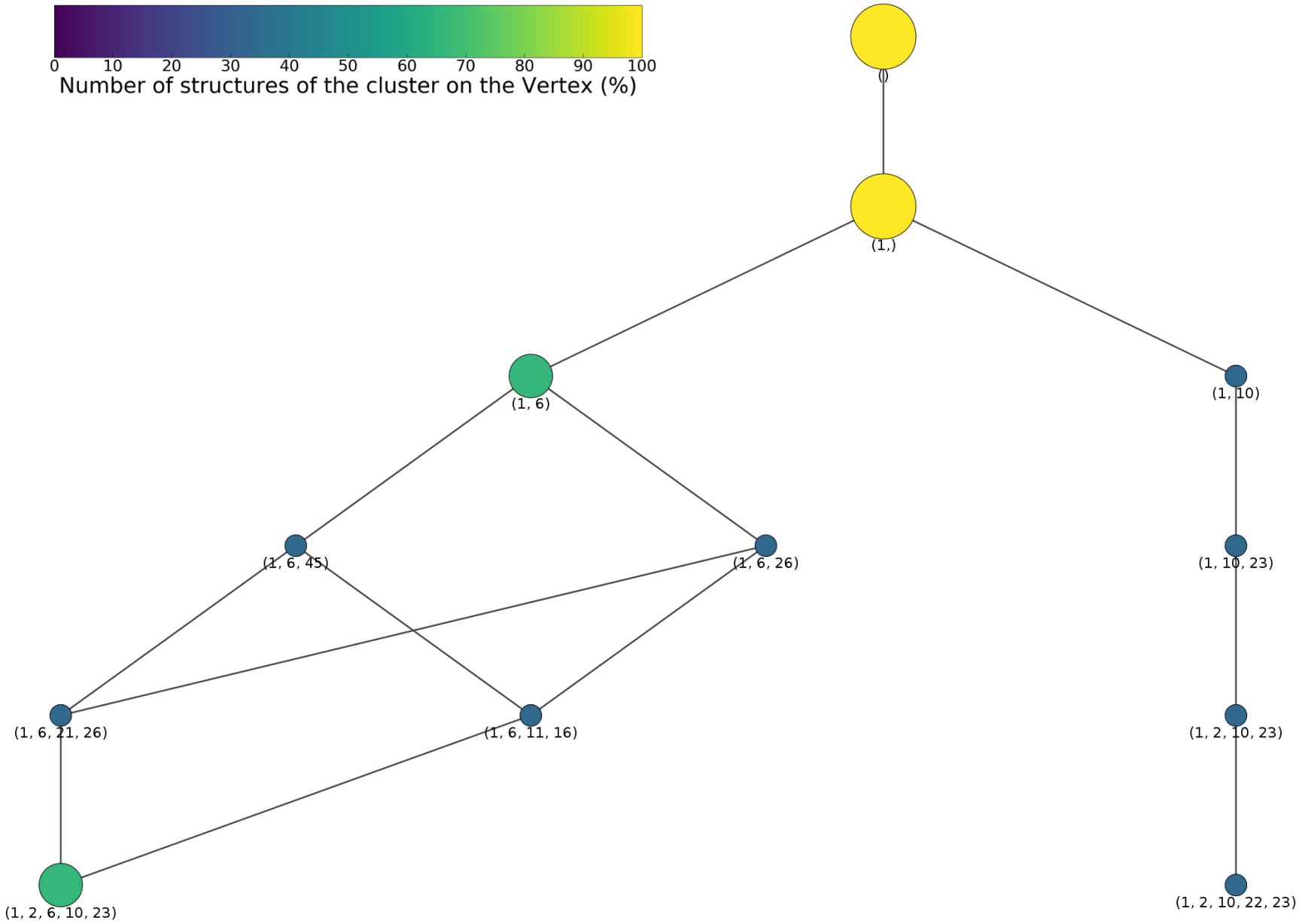

Figure 3.8: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Hepatitis E Virus group in each step of the removal of one protein, with heuristic I.

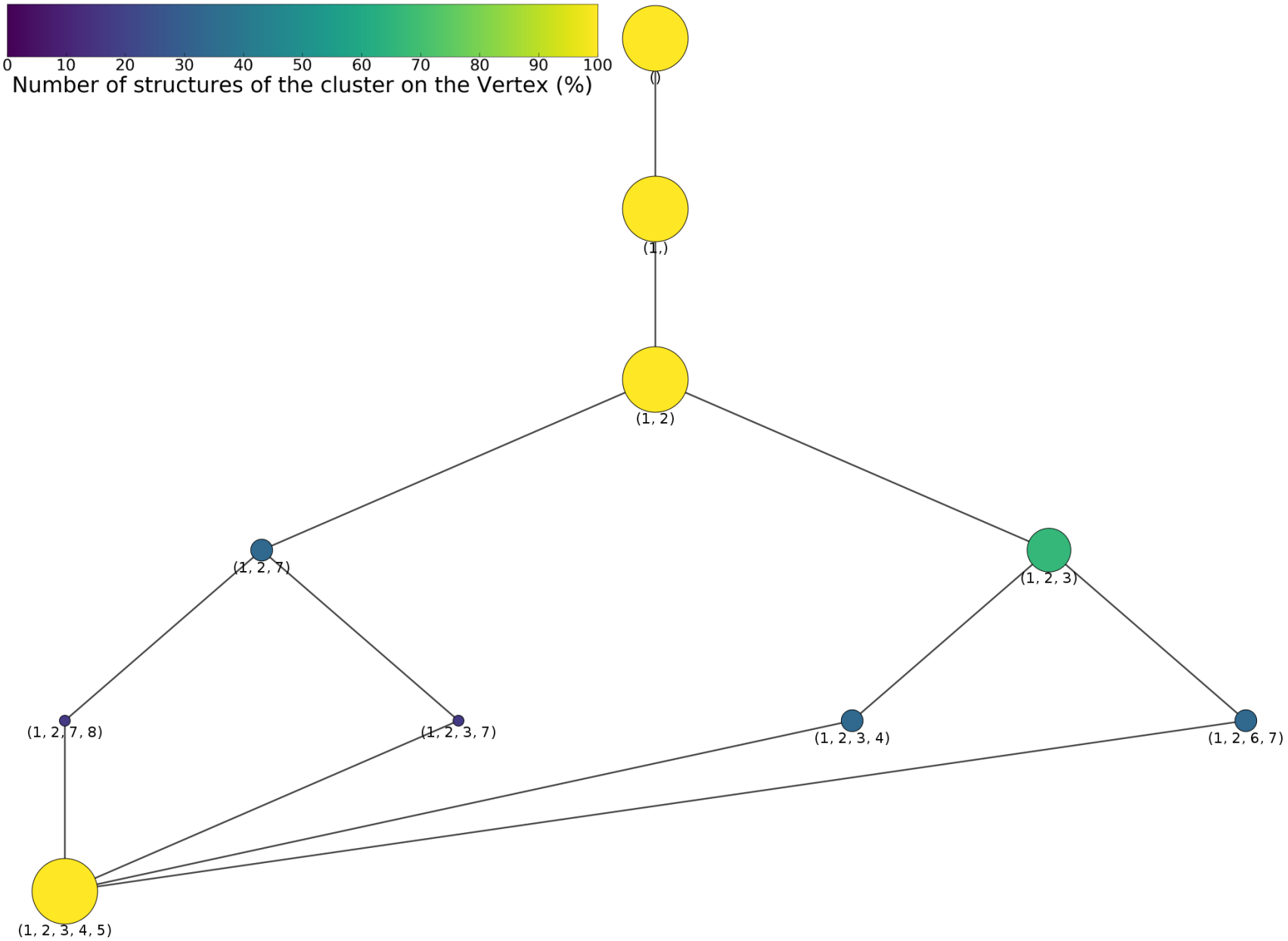

Figure 3.9: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Human Adenovirus Pt-Dd group in each step of the removal of one protein, with heuristic I.

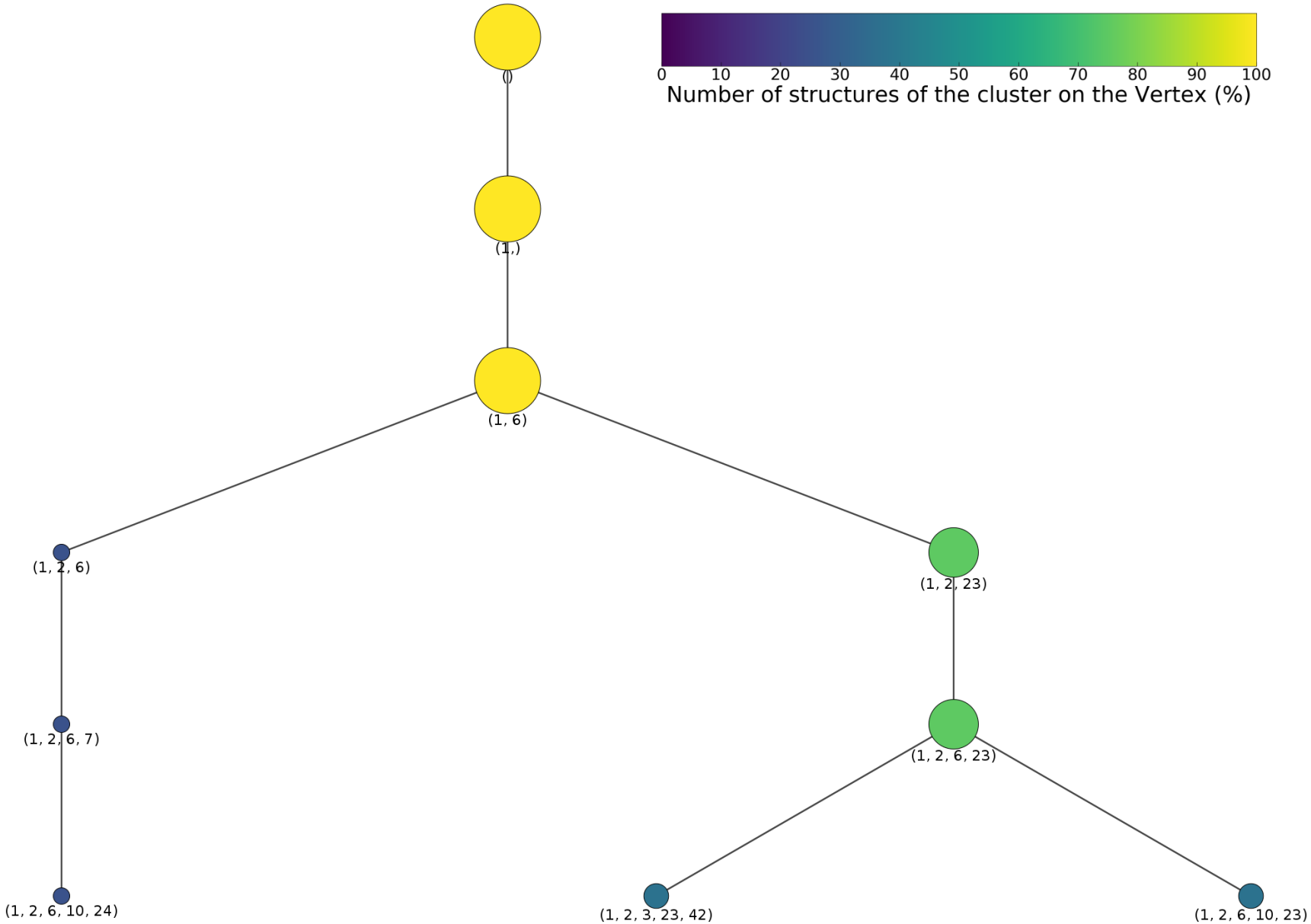

Figure 3.10: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Satellite Tobacco Mosaic Virus group in each step of the removal of one protein, with heuristic I.

Most of the viral capsids analysed tended to end up into two structures: $\{1,2,10,22,23\}$ (Figure 3.16) and $\{1,2,6,10,23\}$ (Figure 3.17). Both involved the removal of the triangle $\{1,10,23\}$, easily seen in the close up of the hole created by the removal of the subunits (Figures 3.16.C and 3.17.C), although, symmetrically, both are very distinct (Figure 3.12.A and Figure 3.12.B). The combination $\{1,2,10,22,23\}$ has a centre on the right-trapezium-like structure made up by the sum of the triangle $\{1,10,23\}$ and the square $\{1,2,22,23\}$ (Figure 3.12.A). On the other hand, the combination $\{1,2,6,10,23\}$ has a centre on the triangle $\{1,10,23\}$, in which the proteins on the right side of two of the vertices of this triangle are removed (Figure 3.12.B), making a structure similar to a scalene trapezium.

Adeno-Associated Virus (Figure 3.1) had a *PDB* structure (PDB_{ID} $2g8g$) that adopted two energetically identical states when 5 proteins were removed, structurally shown on Figures 3.18 and 3.19. Both involved the removal of proteins $\{1,6\}$ and of a triangle on the capsid, that can be pointing to the opposite direction or onto the $\{1,6\}$ proteins ((Figure 3.12.C and D, respectively).

As an exception to the rule, the Human Adenovirus Pt-Dd group (Figure 3.9) followed a five-fold removal of the proteins, taking a pentagon of the capsid (Figure 3.12.E). On Figure 3.20 we can observe the structure of a Human Adenovirus, which is formed by very condensed clusters of pentagons, having very few contacts with the 2-fold and 3-fold proteins.

Satellite Tobacco Mosaic Virus results will not be analysed furthermore since, due to the disparity of the results, the size of the set is too small to draw any reliable conclusions.

3.3 Effect of the Energy Heuristics

The simulations were run again on the same list of structures, this time using heuristics II and III for the calculation of the total energy of the different combinations. The results using heuristic II are shown on Appendix D (Tables D.1 to D.10 and Figures D.1 to D.7), and those using heuristic III are on Appendix E (Tables E.1 to E.10 and Figures E.1 to E.8).

The Adeno-Associated Virus group, under heuristic II, narrowed down the path to a linear one (Figure D.1), compared to the branched one obtained with heuristic I,

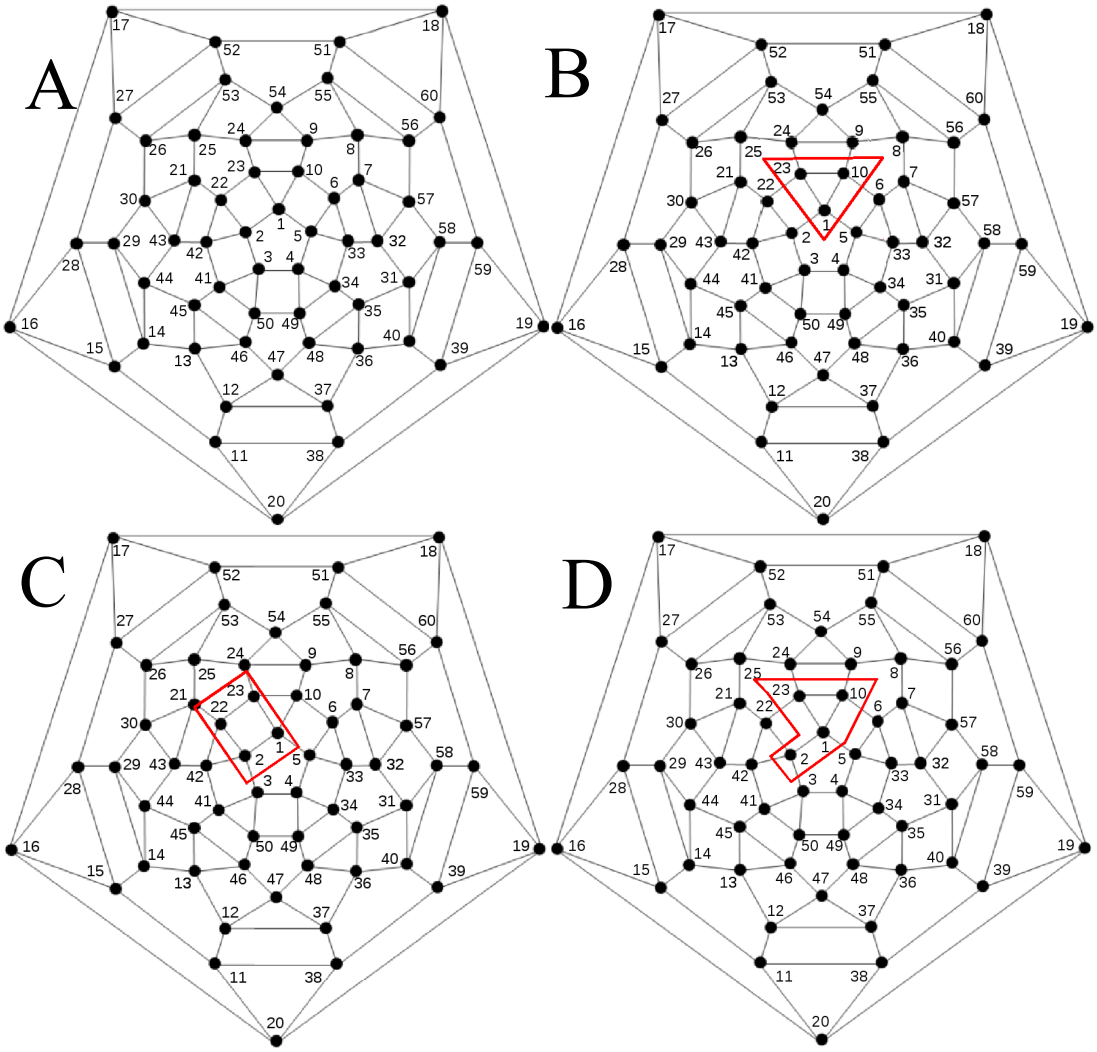

Figure 3.11: Graphs of the Rhombicosadodecahedron (A), indicating the combinations in which the following subunits are removed: $\{1,10,23\}$ (B), $\{1,2,22,23\}$ (C), $\{1,2,10,23\}$ (D).

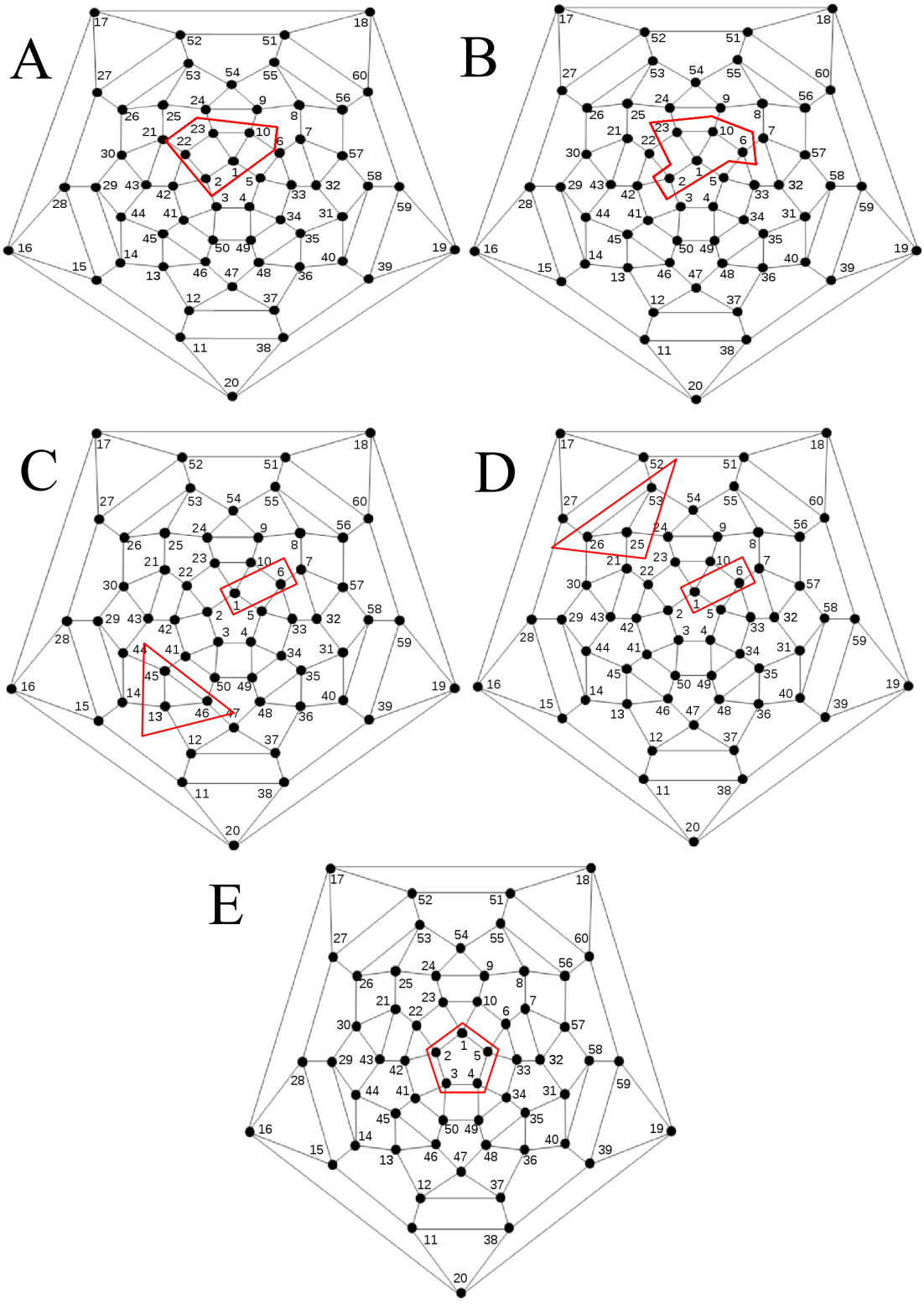

Figure 3.12: Graphs of the Rhombicosadodecahedron, indicating the combinations in which the following subunits are removed: $\{1,2,10,22,23\}$ (A), $\{1,2,6,10,23\}$ (B), $\{1,6,13,45,46\}$ (C), $\{1,6,25,26,53\}$ (D), $\{1,2,3,4,5\}$ (E).

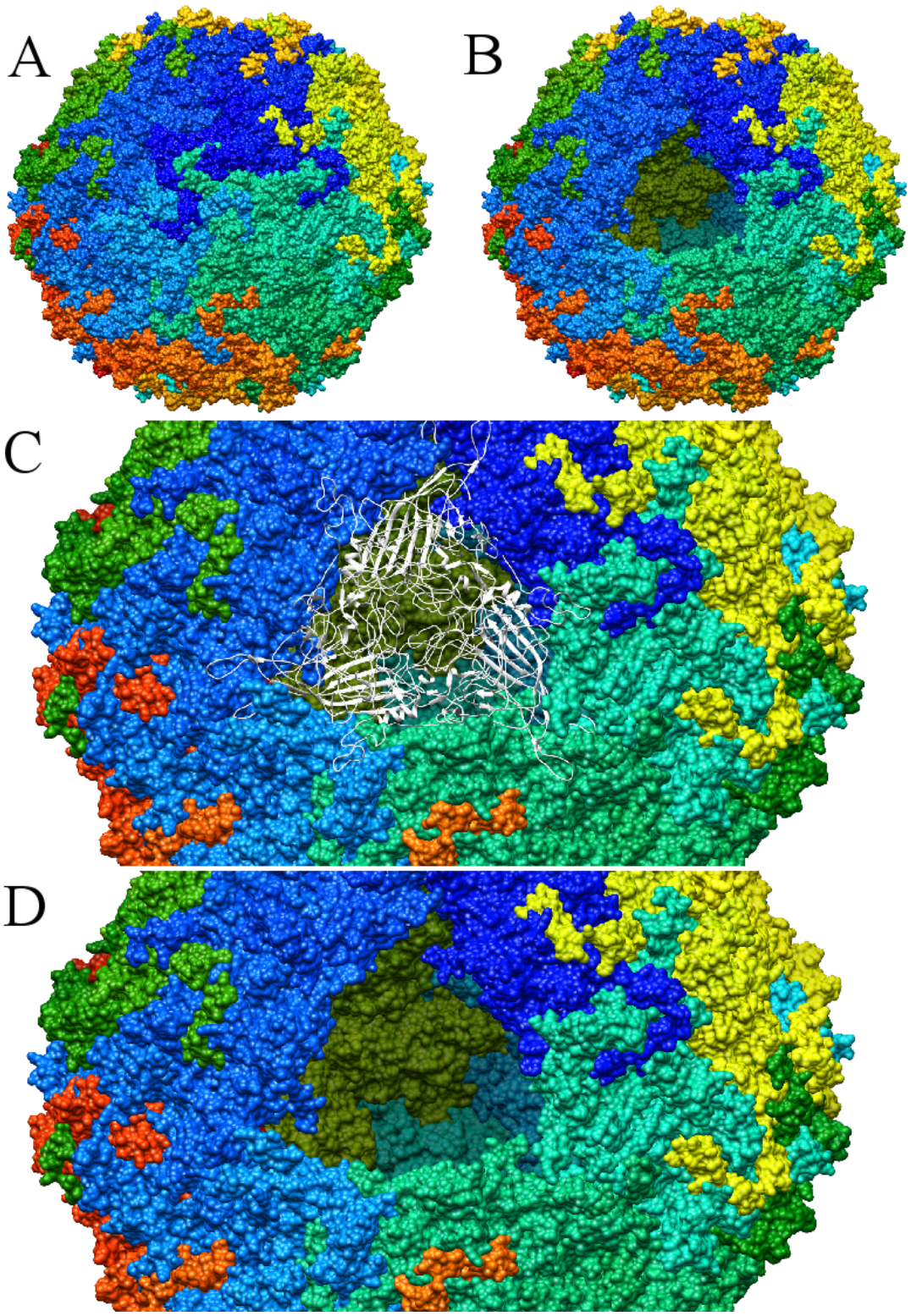

Figure 3.13: Structure of 1s58 (A) by the removal of proteins $\{1, 10, 23\}$ (B, close up: D), with the respective proteins coloured white (C). Structure painted by protein, from blue (first protein) to red (60^{th} protein) in a rainbow pattern.

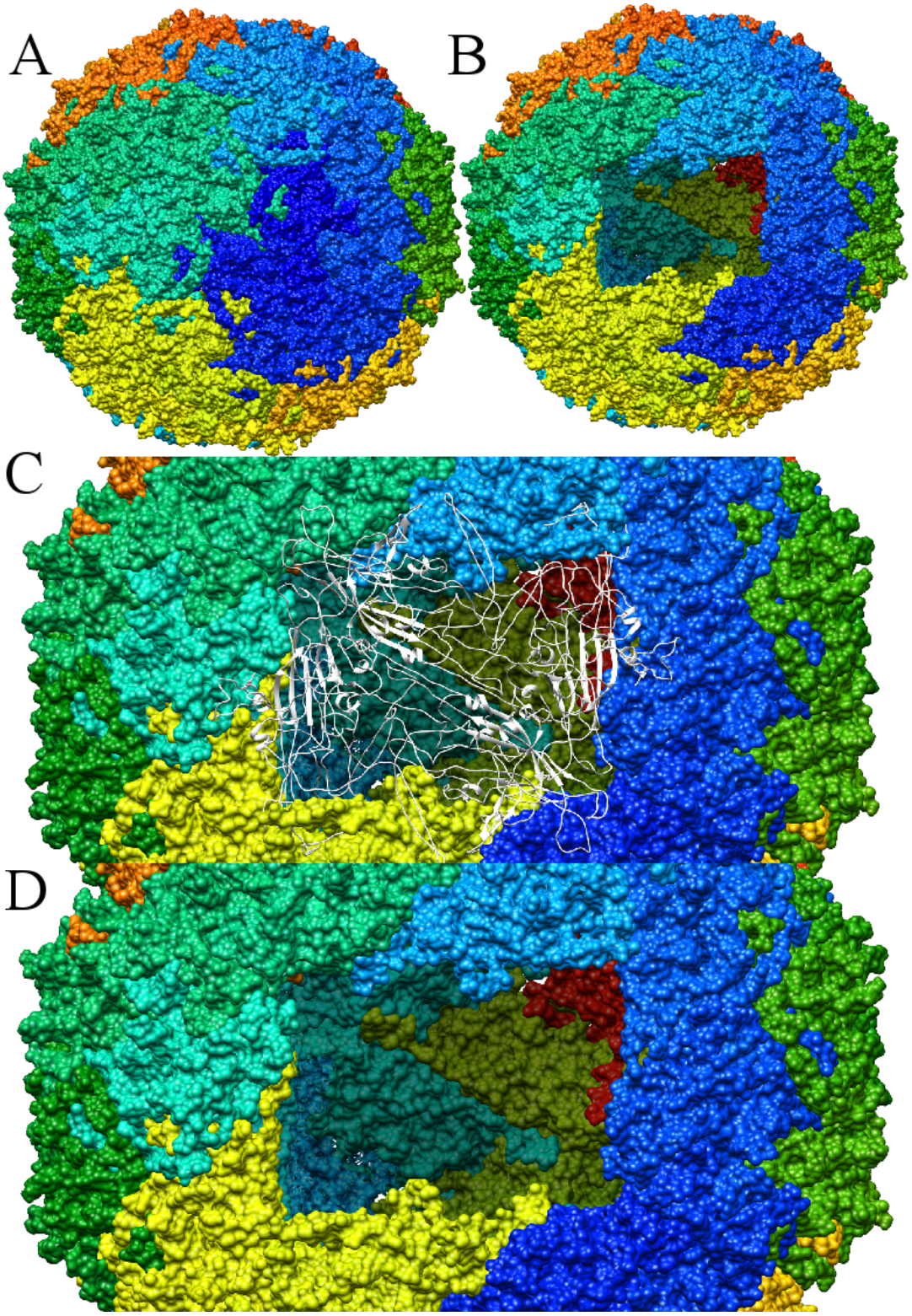

Figure 3.14: Structure of 1dnv (A) by the removal of proteins $\{1, 2, 22, 23\}$ (B, close up: D), with the respective proteins coloured white (C). Structure painted by protein, from blue (first protein) to red (60^{th} protein) in a rainbow pattern.

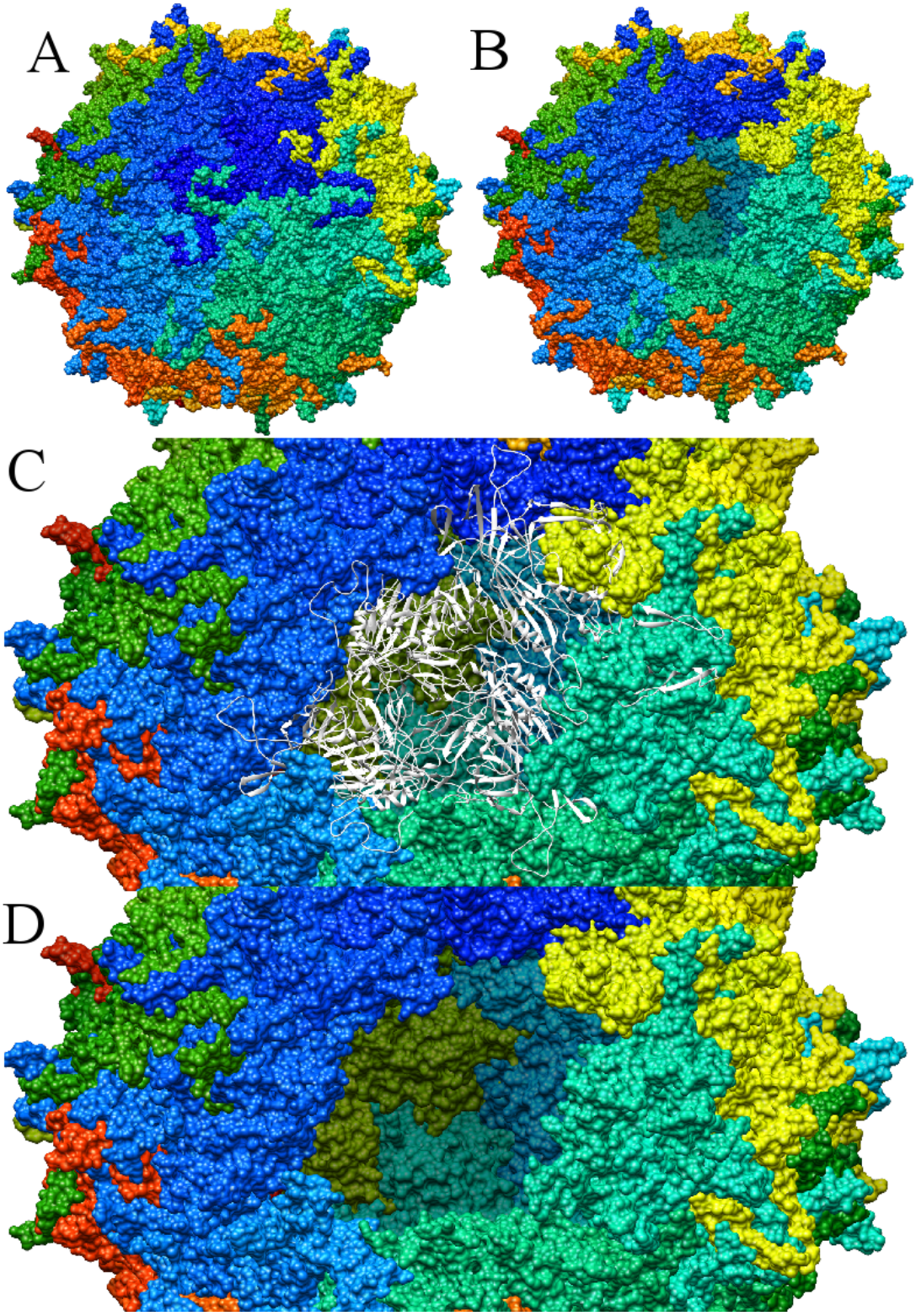

Figure 3.15: Structure of 3raa (A) by the removal of proteins $\{1,2,10,23\}$ (B, close up: D), with the respective proteins coloured white (C). Structure painted by protein, from blue (first protein) to red (60^{th} protein) in a rainbow pattern.

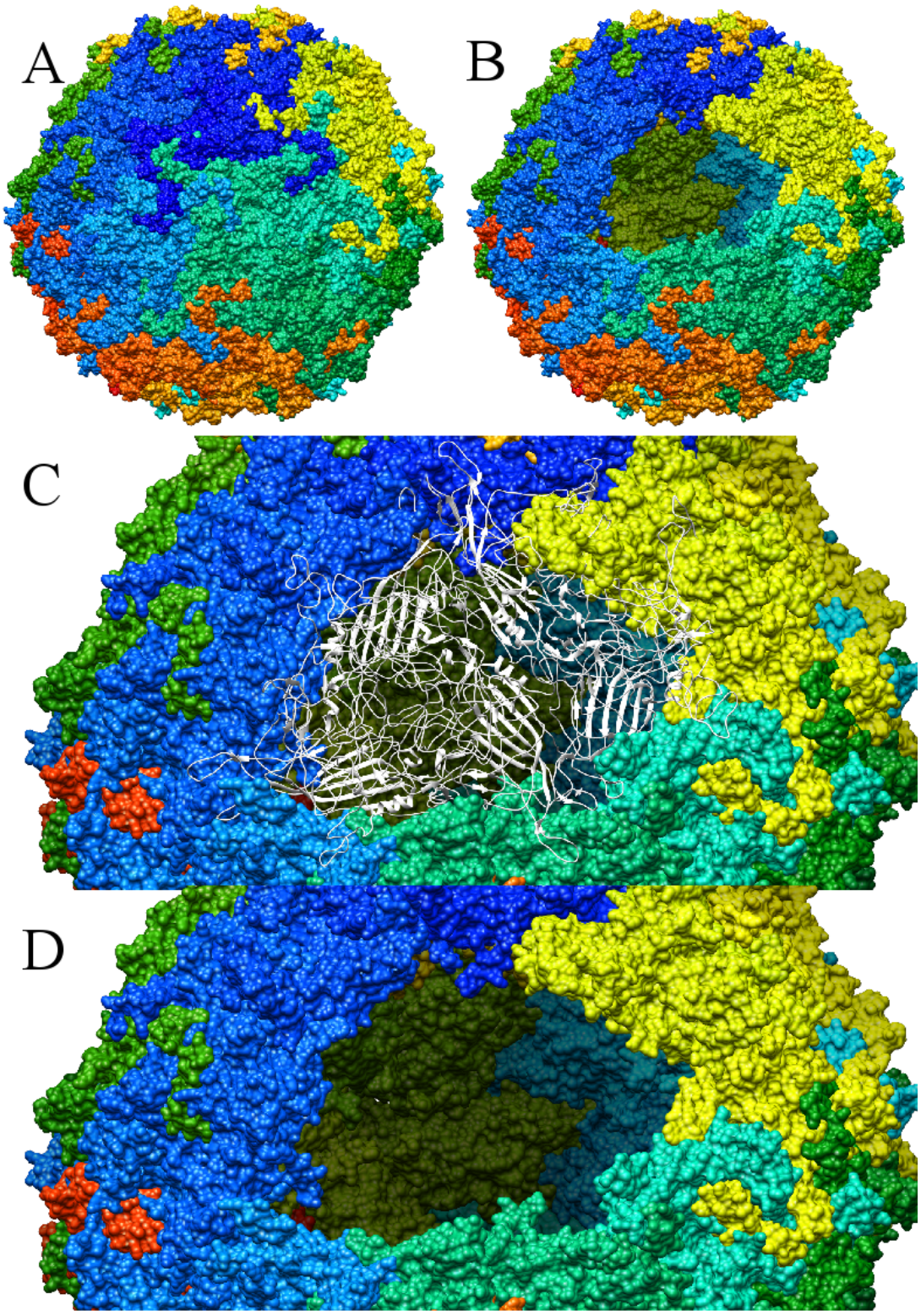

Figure 3.16: Structure of 1s58 (A) by the removal of proteins $\{1, 2, 10, 22, 23\}$ (B, close up: D), with the respective proteins coloured white (C). Structure painted by protein, from blue (first protein) to red (60^{th} protein) in a rainbow pattern.

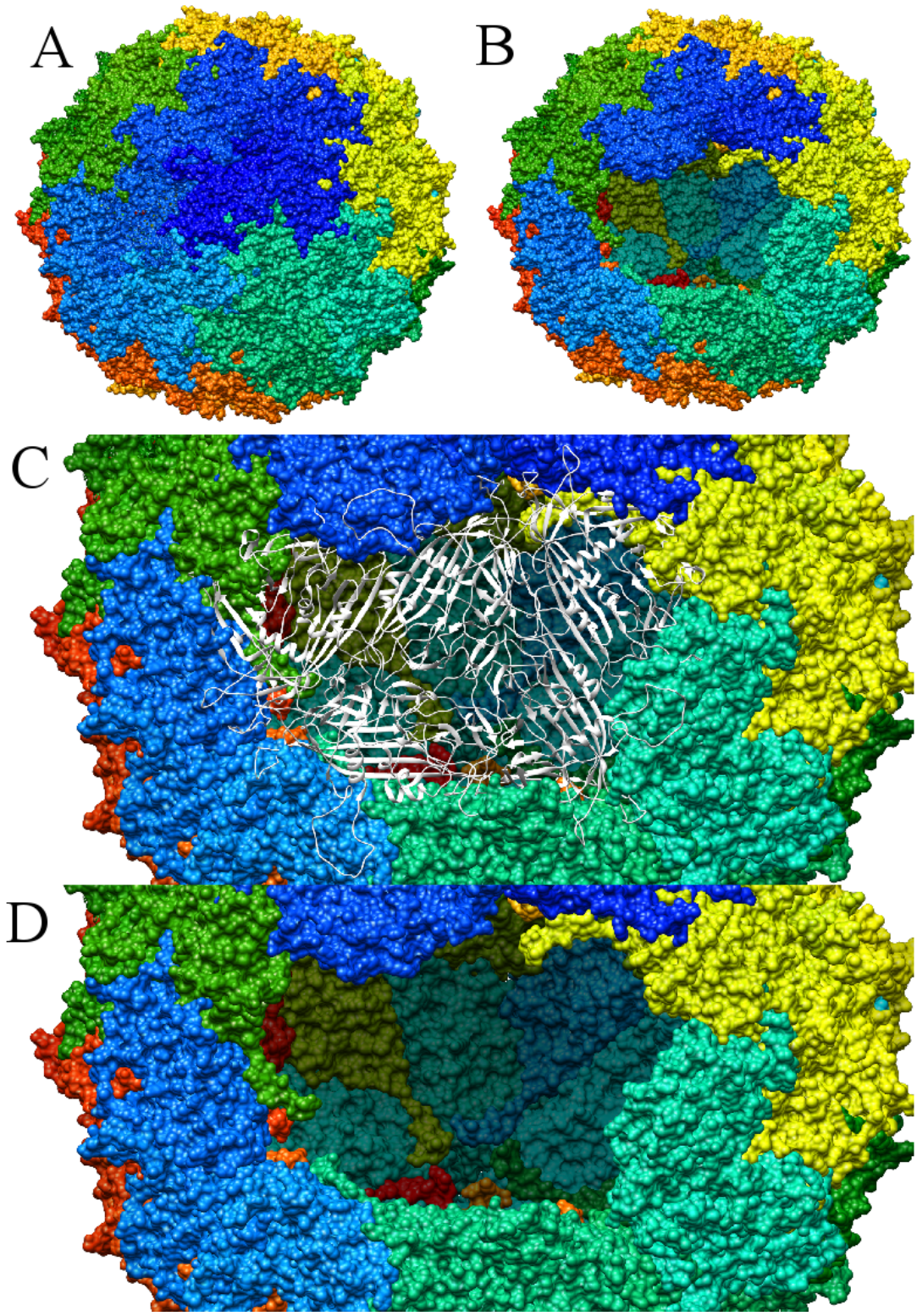

Figure 3.17: Structure of 3p0s (A) by the removal of proteins $\{1, 2, 6, 10, 23\}$ (B, close up: D), with the respective proteins coloured white (C). Structure painted by protein, from blue (first protein) to red (60^{th} protein) in a rainbow pattern.

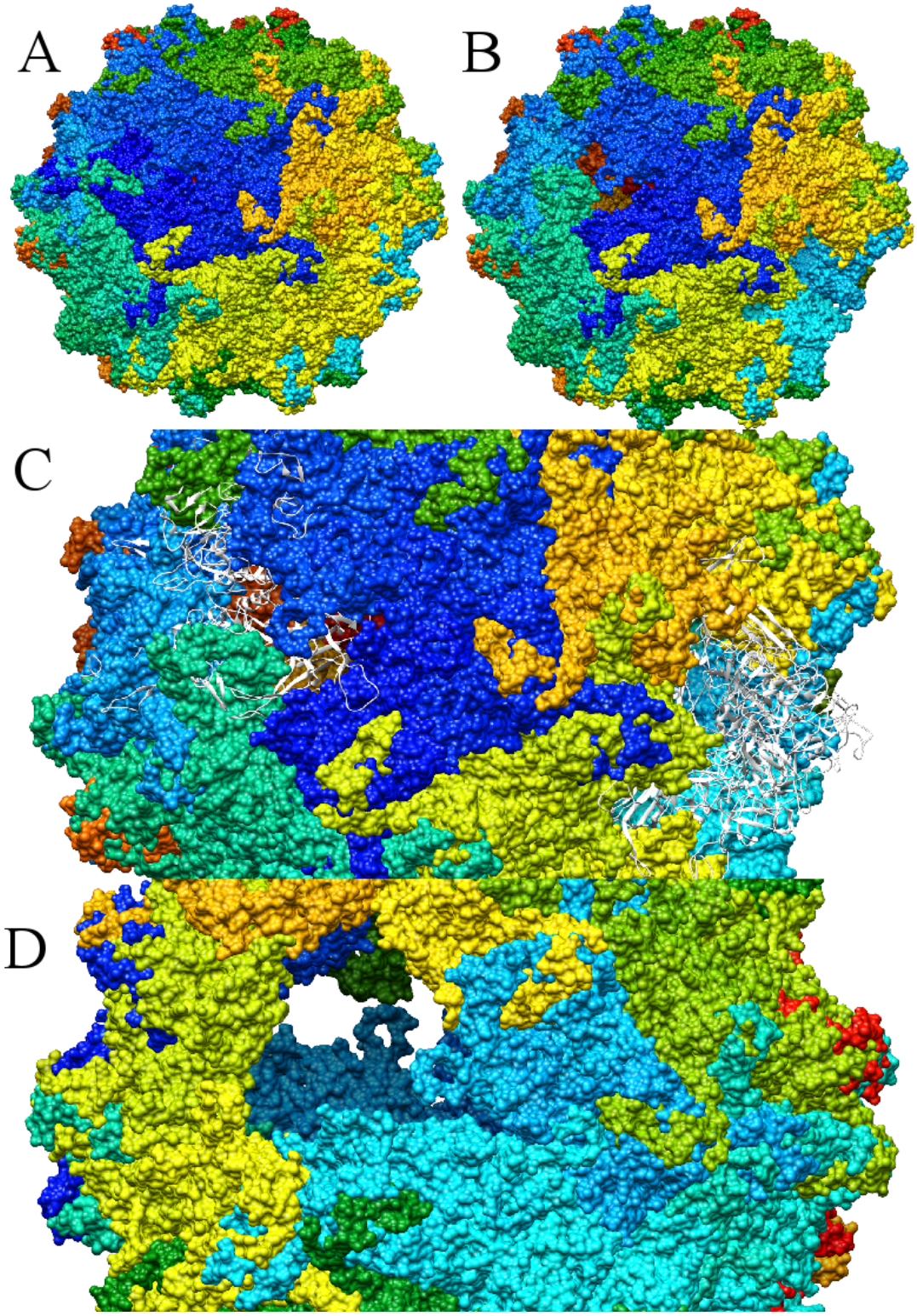

Figure 3.18: Structure of 2g8g (A) by the removal of proteins $\{1,6,13,45,46\}$ (B, close up from another perspective: D), with the respective proteins coloured white (C). Structure painted by protein, from blue (first protein) to red (60^{th} protein) in a rainbow pattern.

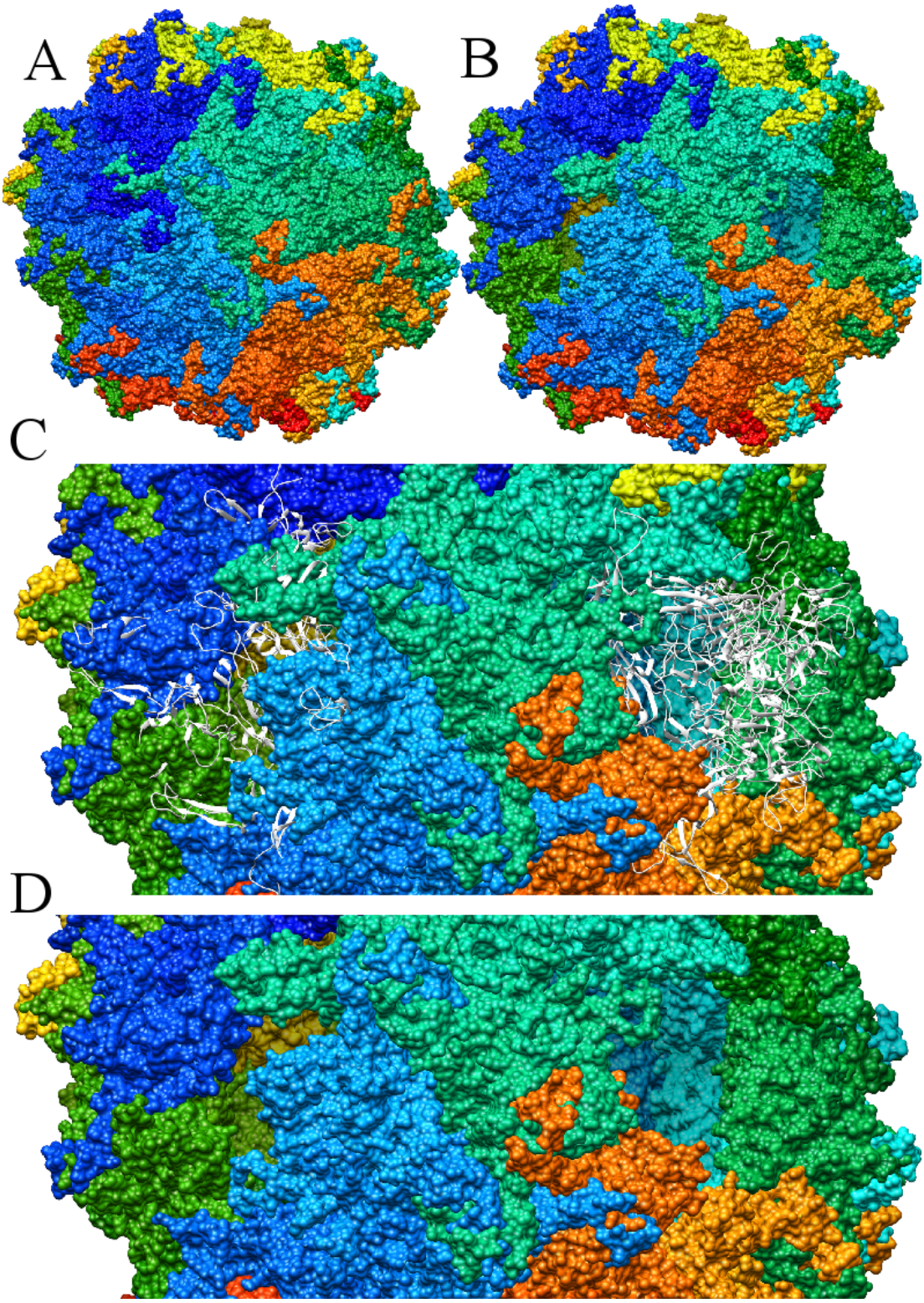

Figure 3.19: Structure of 2g8g (A) by the removal of proteins $\{1,6,25,26,53\}$ (B, close up: D), with the respective proteins coloured white (C). Structure painted by protein, from blue (first protein) to red (60^{th} protein) in a rainbow pattern.

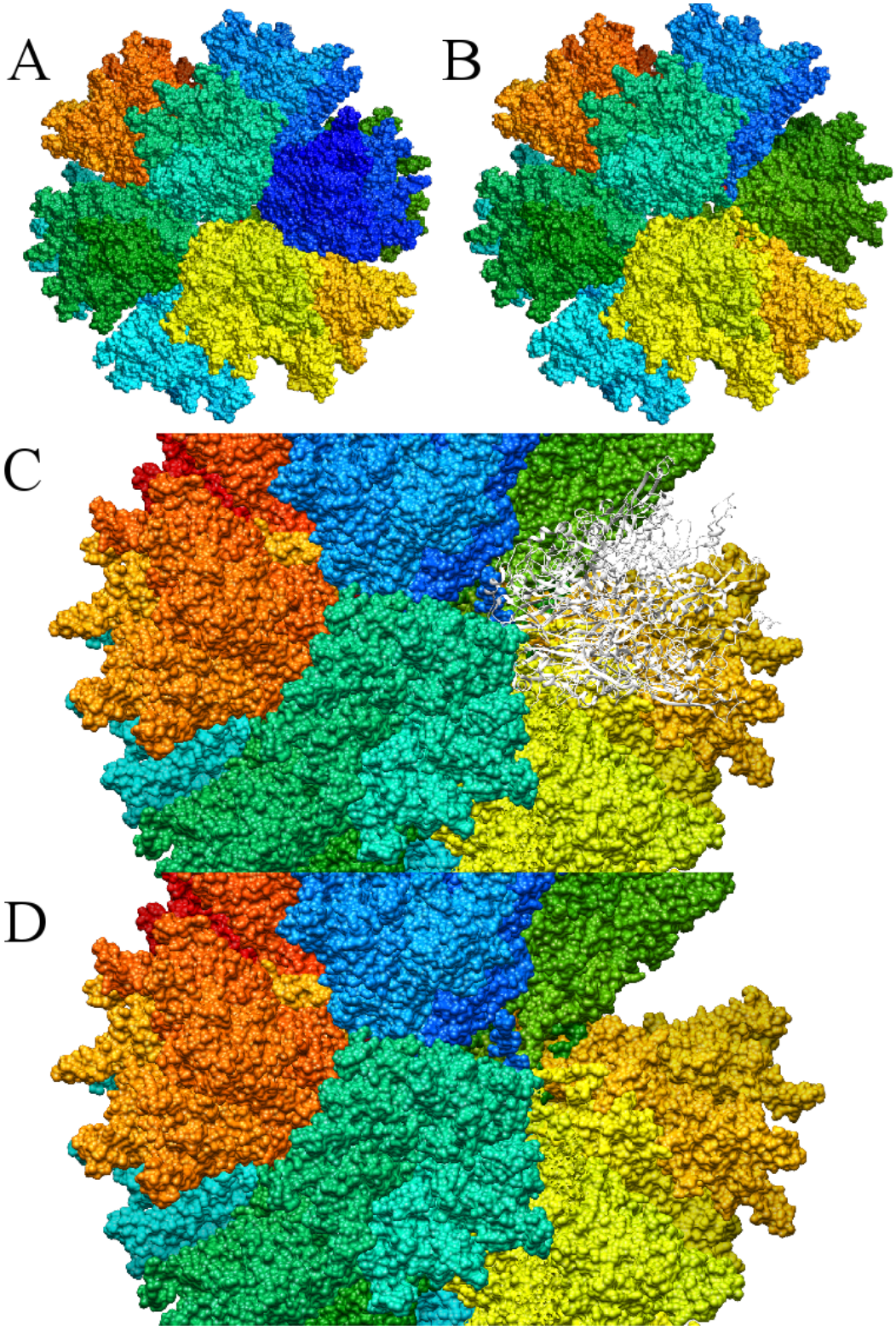

Figure 3.20: Structure of 4aqq (A) by the removal of proteins $\{1,2,3,4,5\}$ (B, close up: D), with the respective proteins coloured white (C). Structure painted by protein, from blue (first protein) to red (60^{th} protein) in a rainbow pattern.

ending up on structure $\{1,2,10,22,23\}$. As for heuristic III, the results (Figure E.1) were identical to the ones of heuristic I (Figure 3.1), except for more branching of the pathways. Comparing Tables C.1 and E.1, structures 3j1q and 3ra8 ended up with a final combination of $\{1,2,6,10,23\}$ on heuristic III, when on heuristic I both were removing proteins $\{1,2,10,22,23\}$ as the last combination.

Avian Birnavirus, Bovine Parvovirus, Human Parvovirus, Porcine Parvovirus and Rodent Protoparvovirus groups, for both heuristics II and III, have shown identical results (Figures D.1 and E.3), following the exact same path. Compared to the results of heuristic I, the only differences to point out are on Avian Birnavirus and Porcine Parvovirus groups, where on the first there was a reduction of the branching of the pathway, from heuristic I to both II and III, turning into a linear path (see Figure 3.7 in comparison to Figures D.1 and E.3); and on the second, the combination $\{1,2,23,30\}$ on heuristic I (Figure 3.2.B) was replaced by $\{1,2,10,23\}$ on the other two (Figures D.1 and E.3).

The results of the Canine and Feline Panleukopenia Virus group, under the heuristics II and III, resemble the previous ones for the main path, with the difference that when there was the removal of four proteins. In both this heuristics there were two options, ending up both on $\{1,2,10,22,23\}$. It should be noticed that the alternative path chosen, besides $\{1,2,10,23\}$, on heuristic II and III, involved always the removal of three proteins arranged in a triangular fashion (see Figure 3.11.A for the positions of combinations $\{1,9,24,54\}$ and $\{1,2,22,42\}$, under heuristics II and III). Compared to the results obtained on heuristic I for this group, a decrease on the branching of the path can be noticed when compared to heuristics II and III (see Figure 3.4 in comparison to Figures D.3 and E.4).

For the Porcine Circovirus group, no change can be noticed between the usage of heuristic I and II (see Figure 3.6 in comparison to Figure D.6). For heuristic III, $\{1,2,22,42\}$ turns out to be the most energetically favourable combination for removing four proteins, and also there is a split on the last stage of the path into two distinct combinations, $\{1,7,23,32,57\}$ and $\{1,2,10,22,23\}$ (Figure E.7). These two minimums, reached by the two structures that build this group, had each other's energy minimum combination as the second energy minimum of the other (Table E.9).

The three Densovirus groups studied (*Bombyx mori*, *Galleria mellonella* and *Penaeus stylirostris* Densoviruses) varied their last stage of the paths under the different

heuristics (as seen on Tables C.3, D.3 and E.3 and on Figures 3.3, D.2 and E.2). Nevertheless, the passage through a stage with the removal of proteins $\{1,2,22,23\}$ is almost consensual, except for the *Bombyx mori* Densivirus group, in which this same stage under the heuristic III was substituted by $\{1,2,6,23\}$ (Figure E.2.A and Table E.3). Even so, for this same group, the second energy minimum for the removal of four proteins was the combination $\{1,2,22,23\}$.

The main changes seen between heuristic I and the other two on Human Adenovirus Pt-Dd group were the reductions of branching of the pathway for disassembly for the last ones. On heuristic II, the pathway resumed to the removal of the five-fold protein, with no sideways path (Figure D.5), while on heuristic III, a divergent option appeared for the removal of four proteins, ending up on structure $\{1,2,3,4,5\}$ (Figure E.6).

Chapter 4

Discussion

As seen by the mechanical removal of proteins with AFM (12), the Minute Mice Virus (on this work represented by the group of the Rodent Protoparvovirus) tends to start the disassembly process by loosing a triangular block, which is then followed by the removal of the adjacent triangle. This is supported by our results, independently of the heuristic used, since most disassembly pathways lead to the removal of a triangular structure, commonly composed of proteins $\{1, 10, 23\}$ (Figures 3.1 to 3.8, D.1 to D.3, D.6, E.1, E.3 to E.5, and E.7). Other structures, besides the removal of the triangle, continue by loosing adjacent proteins, such as the combinations $\{1, 2, 10, 22, 23\}$ and $\{1, 2, 6, 10, 23\}$ (Figures 3.1 to 3.8, 3.10, D.1 to D.4, D.6 to D.7, E.1 to E.5, and E.7 to E.8).

Even outlier cases, such as $\{1, 6, 25, 26, 53\}$ and $\{1, 6, 13, 45, 46\}$ (Figures 3.1 and E.1), loose a triangular structure (Figure 3.12.C and D). This configurations were obtained with structure with PDB_{ID} 2g8g (an Adeno-Associated Virus-4) under heuristics I and III (Tables C.1 and E.1) and, compared with Adeno-Associated Virus-2 (such as structure with PDB_{ID} 1lp3), 2g8g has a depression on the two-fold axis (33), which might reduce the interaction between the subunits of the two-fold (such as proteins $\{1, 6\}$), making them more likely to be removed first (Figures 3.1 and E.1). Independently of this, all the previous results support the idea that the most prone proteins to leave the capsid are those which form a triangle.

From our results, we can speculate that the results of Castellanos et al. (12) might be applicable to more families of viruses besides the Rodent Protoparvovirus, such

as those in the Adeno-Associated Virus group, Avian Birnavirus group, Bovine Parvovirus group, Canine and Feline Panleukopenia Virus Cluster, Human Parvovirus group, Porcine Circovirus group and Porcine Parvovirus group. In line with these predictions, the theoretical studies of Rapaport et al. (22; 23) reveal the existence of long-lived transient structures of capsids with just one last triangle of proteins missing to form the complete capsid. In addition, the results of using individual proteins or trimers in capsid assembly are the same (20), which might suggest that the path for disassembly also follows the triangle removal, as supported by our results.

The disassembly of Densoviruses (Figures 3.3, D.2 and E.2) was mainly predicted to follow a path through structure $\{1,2,22,23\}$. This structure, as seen in both Figures 3.11.C and 3.14, has a two-fold symmetry with a square hole on the capsid. We might speculate that the inter-subunit interactions on these viruses are different from the others, such as parvoviruses. To verify this, we took the ratio of the total number of contacts between $[22 \leftrightarrow 23]$[1] to the total number of contacts between $[10 \leftrightarrow 23]$, the only two interactions between capsid proteins that are in contact and that are different between $\{1,2,22,23\}$ and the common parvovirus structure $\{1,2,10,23\}$, respectively.

Table 4.1: Ratio of contacts between proteins $[22 \leftrightarrow 23]$ and $[10 \leftrightarrow 23]$.

Ratio of the number of contacts	Densovirus			Parvovirus	
	1dnv	3n7x	3p0s	1s58	3raa
$\frac{\#[22 \leftrightarrow 23]}{\#[10 \leftrightarrow 23]}$	$\frac{18}{19} \approx 0,95$	$\frac{15}{11} \approx 1,36$	$\frac{23}{28} \approx 0,82$	$\frac{35}{44} \approx 0,80$	$\frac{35}{62} \approx 0,56$

These results are shown on Table 4.1, where it can be observed that Parvoviruses have a lower ratio than Densovirus, suggesting that

$$\frac{\#[22 \leftrightarrow 23]}{\#[10 \leftrightarrow 23]} \gtrapprox 0.82 \Longrightarrow \{1,2,22,23\}$$

$$\frac{\#[22 \leftrightarrow 23]}{\#[10 \leftrightarrow 23]} \lessapprox 0.80 \Longrightarrow \{1,2,10,23\}$$

[1] where $[\alpha \leftrightarrow \beta]$ means the contacts between protein α and β and $\#[\alpha \leftrightarrow \beta]$ the number of such contacts.

This idea is supported also by the fact that, under heuristic III, structure 3p0s (Table E.3) diverges from the usual $\{1,2,22,23\}$, suggesting that its low value of $\frac{\#[22\leftrightarrow 23]}{\#[10\leftrightarrow 23]}$ makes it more prone for following a path with different configurations. A further analysis of these and other ratios could provide some insight into why some structures are oriented into particular configurations and/or pathways. Besides, the removal of a square-shaped group of proteins raises the idea that the higher the symmetry of the structures removed, maybe more prone they are to be removed.

The removal of five proteins resulted into two main possibilities: $\{1,2,10,22,23\}$ and $\{1,2,6,10,23\}$. Careful analysis of the paths undertaken to arrive to configuration $\{1,2,10,22,23\}$, most of the time they involved the removal of the triangle $\{1,10,23\}$, followed by the removal of vertices $\{2,22\}$. Supposing this trend continues, it is not hard to see that there is a small chance of removing the proteins around the five-fold axis. These could be the follow up steps of the disassembly of these capsids, potentially resulting on one without 15 proteins such as $\{1,2,3,4,5,6,10,22,23,33,34,41,42,49,50\}$ (see Figure 3.11.A to know which proteins would be removed). As for $\{1,2,6,10,23\}$, the steps follow up steps of disassembly are much harder to predict. Even so, following the conclusions of Castellanos et al., some capsids under AFM released a pentamer of triangles (12). Castellanos et al.'s results followed the theoretical results of Reddy et al. (20), where the lowest energy minimum configuration, just before the complete capsid, was the one missing a pentamer of triangles (corresponding to the 15-protein combinations described before), inclusively for Parvoviruses (21).

In the Hepatitis E Virus group, it is interesting to underline that, for both heuristics I and III, structure with PDB_{ID} 3hag follows a different path than the other two (Tables C.4 and E.4), which does not happen for heuristic II. It is important also to notice that 3hag has mutations on 4 amino acids (34), whereas structures with PDB_{ID} 2ztn (35) and 2zzq (36) have none and one, respectively. This suggests that a small number of mutations can interfere with the pathways of disassembly.

The Human Adenovirus Pt-Dd group is the only taking a complete different disassembly path, removing the pentamer $\{1,2,3,4,5\}$ (Figures 3.9, D.5 and E.6 and tables C.10, D.10 and E.10), independently of the heuristic used. The distance between the pentagonal clusters on the full capsid structure (Figure 3.20) might make it eas-

ier for this set of subunits to be removed, in opposition to creating a bigger gap by removing, for example, proteins $\{1, 2, 10, 22, 23\}$.

There is no significant difference between the results obtained through the different heuristics. Heuristic II gives an equal weighting to all types of contacts. Interestingly, this heuristic ends up giving very similar results to the other two, suggesting that the variation of the number of contacts, more than their energy, might have a leading role on the determination of the path of disassembly. This heuristic predicted paths of disassembly that were similar to the experimental results of Reddy et al. (20; 21), Horton et al. (24) and Rapaport et al. (22; 23). Moreover, this heuristic is simplest of the three, making it the strongest candidate to be used in further computational studies of the disassembly process.

The Satellite Tobacco Mosaic Virus group was not analysed into any depth since the predictions subdivided into almost as many different paths as the number of capsid structures that make up the group. To have more informative results, the size of the group should be increased in subsequent studies.

Chapter 5

Conclusion

Our study started out with the goal to ascertain whether there was a common disassembly pathway among different virus families and, if not, if there would be a specific conserved disassembly path for each family that could lead into predicting the family based on the observed path of disassembly.

The results on this work support the idea that, for a large cluster of $T = 1$ viruses, there is a common disassembly pathway. This cluster is composed by some Parvoviruses (Adeno-Associated Virus, Bovine Parvovirus, Human Parvovirus, Porcine Parvovirus, Rodent Protoparvovirus, Canine and Feline Panleukopenia Virus), Avian Birnavirus, Hepatitis E Virus and Porcine Circovirus. Densoviruses (from *Bombyx mori*, *Galleria mellonella* and *Penaeus stylirostris*) are exceptional for their exclusive removal of proteins $\{1, 2, 22, 23\}$, with no branching to alternative possibilities. It is also important to see that mutations can affect our results, seen on the different results obtained on the Hepatitis E Virus group, being important since it may affect the way viruses disassemble.

On the other hand, Human Adenoviruses Pt-Dd have a very particular pathway of disassembling, establishing that not all disassemblies are done through the removal of triangular trimers of proteins, but also can be by the removal of pentagonal pentamers of proteins. Nevertheless, it is an error to admit that any $T = 1$ capsid with such a disassembly must be a Human Adenoviruses Pt-Dd, since not all $T = 1$ viruses were studied on this work and many viruses' families presented the same disassembly path on the previous results of this work.

Another aim of this work was the study of different heuristical measures of the total energy. Overall, the results showed no differences under the three heuristics proposed. Moreover, by using a heuristic only based on the number of contacts and not on the type of chemical bonds (heuristic II), we can achieve the same results with less branching from a main pathway.

Using a combinatorial method based on symmetry and geometry, with a 60-subunit model (Deltoidal Hexecontahedron) and not a 20-subunit model (Icosahedron) or 12-subunit model (Dodecahedron), under a heuristic using the number of weak contacts, we obtained comparable results to those found in previous literature. Furthermore, the usage of a 60-subunit model allows the study of cases such as those of Human Adenoviruses, which do not loose triangular faces of the Icosahedron, but pentagon faces of the Deltoidal Hexecontahedron. Nonetheless, an increase of the sample size is an important follow-up step, since it would give a better insight into the trends observed for the different viruses' studies. Moreover, the number of subunits removed from the capsid structures should be above 5 in order to offer a perspective of what would be the next steps of the disassembly pathway. This was not done on this dissertation due to lack of computational power to increase over the 5 subunits. Nevertheless, an improvement of the algorithms as well as the processing power could provide more informative results.

The understanding of the way viruses' capsids disassemble can be used to interfere with the inter-subunit contacts, specially those who hold the triangular trimer to the complete capsid. Antiviral drugs can be design to affect these same contacts, by decreasing or increasing the strength of such contacts, inducing premature disassembly or making it more difficult, respectively, reducing the chance of infection of the host.

References

[1] E. V. Koonin, T. G. Senkevich, and V. V. Dolja, "The ancient Virus World and evolution of cells," *Biol. Direct*, vol. 1, p. 29, 2006. 1

[2] M. M. Poranen, R. Daugelavičius, and D. H. Bamford, "Common Principles in Viral Entry," *Annual Review of Microbiology*, vol. 56, pp. 521–538, Oct. 2002. 1, 2

[3] M. Suomalainen and U. F. Greber, "Uncoating of non-enveloped viruses," *Current Opinion in Virology*, vol. 3, pp. 27–33, Feb. 2013. 1, 2

[4] Y. Yamauchi and A. Helenius, "Virus entry at a glance," *Journal of Cell Science*, vol. 126, pp. 1289–1295, Mar. 2013. 1, 2

[5] D. L. D. Caspar and A. Klug, "Physical Principles in the Construction of Regular Viruses," *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 27, pp. 1–24, Jan. 1962. 1, 2, 3

[6] B. V. V. Prasad and M. F. Schmid, "Principles of Virus Structural Organization," in *Viral Molecular Machines* (M. G. Rossmann and V. B. Rao, eds.), vol. 726, pp. 17–47, Boston, MA: Springer US, 2012. 1, 2, 3

[7] J. Overbaugh, A. D. Miller, and M. V. Eiden, "Receptors and Entry Cofactors for Retroviruses Include Single and Multiple Transmembrane-Spanning Proteins as well as Newly Described Glycophosphatidylinositol-Anchored and Secreted Proteins," *Microbiology and Molecular Biology Reviews*, vol. 65, pp. 371–389, Sept. 2001. 1

[8] M. G. Rossmann, "Viral cell recognition and entry," *Protein Science*, vol. 3, pp. 1712–1725, Oct. 1994. 1

[9] G. R. Nemerow, "Cell Receptors Involved in Adenovirus Entry," *Virology*, vol. 274, pp. 1–4, Aug. 2000. 1

[10] G. R. Whittaker, M. Kann, and A. Helenius, "Viral entry into the nucleus," *Annu. Rev. Cell Dev. Biol.*, vol. 16, pp. 627–651, 2000. 2

[11] "World Health Organization." `https://www.who.int/`. Accessed: 2016-09-16. 2

[12] M. Castellanos, R. Pérez, P. Carrillo, P. de Pablo, and M. Mateu, "Mechanical Disassembly of Single Virus Particles Reveals Kinetic Intermediates Predicted by Theory," *Biophysical Journal*, vol. 102, pp. 2615–2624, June 2012. 2, 6, 43, 45

[13] J. D. Perlmutter and M. F. Hagan, "Mechanisms of Virus Assembly," *Annual Review of Physical Chemistry*, vol. 66, pp. 217–239, Apr. 2015. 2, 3

[14] A. Klug and D. Caspar, "The Structure of Small Viruses," in *Advances in Virus Research*, vol. 7, pp. 225–325, Elsevier, 1961. xv, 3

[15] M. G. Mateu, "Assembly, stability and dynamics of virus capsids," *Archives of Biochemistry and Biophysics*, vol. 531, pp. 65–79, Mar. 2013. 3, 4

[16] A. Zlotnick, "To build a virus capsid. An equilibrium model of the self assembly of polyhedral protein complexes," *J. Mol. Biol.*, vol. 241, pp. 59–67, Aug. 1994. 3, 5, 6

[17] A. Zlotnick, J. M. Johnson, P. W. Wingfield, S. J. Stahl, and D. Endres, "A theoretical model successfully identifies features of hepatitis B virus capsid assembly," *Biochemistry*, vol. 38, pp. 14644–14652, Nov. 1999. xv, 3, 5

[18] A. Zlotnick, "Are weak protein–protein interactions the general rule in capsid assembly?," *Virology*, vol. 315, pp. 269–274, Oct. 2003. 3

[19] A. Zlotnick and S. J. Stray, "How does your virus grow? Understanding and interfering with virus assembly," *Trends Biotechnol.*, vol. 21, pp. 536–542, Dec. 2003. 3, 4, 5, 6

[20] V. S. Reddy, H. A. Giesing, R. T. Morton, A. Kumar, C. B. Post, C. L. Brooks, and J. E. Johnson, "Energetics of Quasiequivalence: Computational Analysis of Protein-Protein Interactions in Icosahedral Viruses," *Biophysical Journal*, vol. 74, pp. 546–558, Jan. 1998. 5, 6, 44, 45, 46

[21] V. S. Reddy and J. E. Johnson, "Structure-Derived Insights into Virus Assembly," in *Advances in Virus Research*, vol. 64, pp. 45–68, Elsevier, 2005. 5, 6, 45, 46

[22] D. C. Rapaport, "Role of Reversibility in Viral Capsid Growth: A Paradigm for Self-Assembly," *Physical Review Letters*, vol. 101, Oct. 2008. 5, 6, 44, 46

[23] D. C. Rapaport, "Studies of reversible capsid shell growth," *Journal of Physics: Condensed Matter*, vol. 22, p. 104115, Mar. 2010. 5, 6, 44, 46

[24] N. Horton and M. Lewis, "Calculation of the free energy of association for protein complexes," *Protein Science*, vol. 1, pp. 169–181, Dec. 2008. 5, 6, 46

[25] Wikipedia, the free encyclopedia, "Deltoidal hexecontahedron," 2016. [Online; accessed April 25, 2016]. xv, 9

[26] W. Burnside, "Theory of groups of finite order," *Messenger of Mathematics*, vol. 23, p. 112, 1909. 8

[27] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000. 10

[28] M. Carrillo-Tripp, C. M. Shepherd, I. A. Borelli, S. Venkataraman, G. Lander, P. Natarajan, J. E. Johnson, C. L. Brooks, and V. S. Reddy, "Viperdb2: an enhanced and web api enabled relational database for structural virology," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D436–D442, 2009. 10

[29] P. W. Atkins and J. De Paula, *Physical chemistry for the life sciences*. Oxford, UK : New York: Oxford University Press ; W.H. Freeman, 2006. 12

[30] "Anaconda software distribution." Computer software. Vers. 2-2.4.0. Continuum Analytics, Nov. 2016. Web. `https://continuum.io`. 16

[31] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal*, vol. Complex Systems, p. 1695, 2006. 16

[32] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "Ucsf chimera—a visualization system for exploratory research and analysis," *Journal of computational chemistry*, vol. 25, no. 13, pp. 1605–1612, 2004. 16

[33] L. Govindasamy, E. Padron, R. McKenna, N. Muzyczka, N. Kaludov, J. A. Chiorini, and M. Agbandje-McKenna, "Structurally Mapping the Diverse Phenotype of Adeno-Associated Virus Serotype 4," *Journal of Virology*, vol. 80, pp. 11556–11570, Dec. 2006. 43

[34] T. S. Y. Guu, Z. Liu, Q. Ye, D. A. Mata, K. Li, C. Yin, J. Zhang, and Y. J. Tao, "Structure of the hepatitis E virus-like particle suggests mechanisms for virus assembly and receptor binding," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 12992–12997, Aug. 2009. 45

[35] T. Yamashita, Y. Mori, N. Miyazaki, R. H. Cheng, M. Yoshimura, H. Unno, R. Shima, K. Moriishi, T. Tsukihara, T. C. Li, N. Takeda, T. Miyamura, and Y. Matsuura, "Biological and immunological characteristics of hepatitis E virus-like particles based on the crystal structure," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 12986–12991, Aug. 2009. 45

[36] L. Xing, T.-C. Li, N. Mayazaki, M. N. Simon, J. S. Wall, M. Moore, C.-Y. Wang, N. Takeda, T. Wakita, T. Miyamura, and R. H. Cheng, "Structure of Hepatitis E Virion-sized Particle Reveals an RNA-dependent Viral Assembly Pathway," *Journal of Biological Chemistry*, vol. 285, pp. 33175–33183, Oct. 2010. 45

Appendix A

Mathematical Background

A.1 Combinatorics and Groups

A.1.1 Combinations

In mathematics, combinations count the number of ways of selecting k elements from a list of size n, where the order in which we select the elements does not matter. The number of combinations is calculated with the expression in Equation A.1.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{A.1}$$

Some of the properties of combinations are seen very easily in Pascal's Triangle. The three properties presented below will not be proved on this work.

$$\binom{n}{k} = \binom{n}{n-k} \tag{A.2}$$

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k} \tag{A.3}$$

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n \tag{A.4}$$

A.1.1.1 Why do Combinations Fail in our Problem?

Take the example of a square where we have four edges and we want to know how many ways there are to remove one of these edges. Through combinations we know that it should be $\binom{4}{1} = 4$. but taking the symmetry of the square into account, we can easily see that all 4 possibilities that the combinations provide are symmetrically equal (imagine rotating the square by multiples of 90º and we can obtain all the combinations). Therefore, in reality, there is only one possibility of removing an edge. Moreover, it is easy to see that the number of possibilities of removing an edge from any symmetrical polyhedra is always going to be smaller or equal to the number predicted by the combinations.

A.1.2 Symmetry Groups

A group is an algebraic structure built with a set of elements, G, and an operation $\diamond$ that is able to join two elements of the set, g_1 and g_2, to form another element, denoted as $g_1 \diamond g_2$. The four axioms that make the set and the operation $\langle G, \diamond \rangle$ a group are:

Closure: For $g_1, g_2 \in G$, $g_1 \diamond g_2 \ \in G$

Associativity: For g_1, g_2 and $g_3 \in G$, $(g_1 \diamond g_2) \diamond g_3 = g_1 \diamond (g_2 \diamond g_3)$

Identity Element: $\exists!\ e \in G$ such that $\forall\ g \in G,\ g \diamond e = e \diamond g = g$

Inverse Element: $\forall\ g_1 \in G\ \ \exists\ g_{-1} \in G$ such that $g_1 \diamond g_{-1} = g_{-1} \diamond g_1 = e$

If the operation is commutative, i.e. for $g_1, g_2 \in G$, $g_1 \diamond g_2 = g_2 \diamond g_1$, the group is abelian. In the cases studied here (symmetry groups) the groups are not abelian, so it cannot be assumed that the operation is commutative.

A symmetry group is a group where its elements are the symmetric operations on points in the Euclidean space. The possible symmetric elements are:

Identity: E , in which all points stay in the same position they were;

Proper Rotation of Order n: C_n , in which all points are rotated $360/n$ degrees around an axis;

Reflection: σ , in which all points are moved along the axis perpendicular to the reflection plane, to the opposite side, maintaining the same distance to the plane;

Inversion: i , in which all points are moved by the application of two reflections, where their reflection planes are perpendicular;

Improper Rotation of Order n: S_n , in which all points are rotated $360/n$ degrees around an axis and reflected through a plane perpendicular to the rotation axis.

A set with these elements can form a group and different sets form distinct groups. Some of these groups are very common and associated with symmetric polyhedra, being represented by a symbol which describes all the symmetry elements formed by it. Some of these groups are presented in Table A.1 and the symmetry operations are organised by their classes.[1] The symmetry group used in this work is the pure rotation icosahedral group I.

Table A.1: Symmetry classes included in some symmetry groups.

Group	Group Elements divided into each class
C_4	E , C_4 , C_2 , C_4^3
D_5	E , $2C_5$, $2C_5^2$, $5C_2$
C_{4v}	E , $2C_4$, C_2 , $2\sigma_v$, $2\sigma_d$
C_{4h}	E , C_4 , C_2 , C_4^3 , i , S_4^3 , σ_h , S_4
T	E , $4C_3$, $4C_3^2$, $3C_2$
T_h	E , $4C_3$, $4C_3^2$, $3C_2$, i , $4S_6$, $4S_6^5$, $3\sigma_h$
T_d	E , $8C_3$, $3C_2$, $6S_4$, $6\sigma_d$
O	E , $6C_4$, $3C_2(=C_4^2)$, $8C_3$, $6C_2$
O_h	E , $8C_3$, $6C_2$, $6C_4$, $3C_2(=C_4^2)$, i , $6S_4$, $8S_6$, $3\sigma_h$, $6\sigma_d$
I	E , $12C_5$, $12C_5^2$, $20C_3$, $15C_2$
I_h	E , $12C_5$, $12C_5^2$, $20C_3$, $15C_2$, i , $12S_{10}$, $12S_{10}^3$, $20S_6$, 15σ

A.1.3 Cyclical Permutations and Symmetry Elements

A symmetry element, such as a rotation, can be expressed in the form of a permutation of the points in space. Since the symmetry operations are cyclical (i.e. after

[1] Two operations g_α and g_β are from the same class if $\exists\, g_x$ such that $g_x \diamond g_\alpha \diamond g_x^{-1} = g_\beta$ $(g_x \diamond g_x^{-1} = e)$

applying them a number of times, the points end up in the beginning positions), we can write the operation such as cyclical permutations.

To understand this concept, we will use the example of the rotation of a numbered square on a plane. The rotation of the square is a C_4 rotation, having four possible outcomes $\left(C_4,\, C_4^2,\, C_4^3,\, C_4^4 = E\right)$. A simple C_4 rotation is described on Figure A.1.

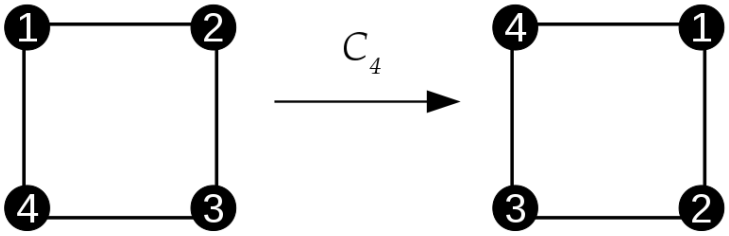

Figure A.1: Vertex-Numbered Square C_4 rotation

A permutation that is equivalent to the rotation would be the one described on Equation A.5, being the cyclical version of this permutation on Equation A.6.

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix} \tag{A.5}$$

$$(1 \to 2,\ 2 \to 3,\ 3 \to 4,\ 4 \to 1) \Longleftrightarrow (1,\ 2,\ 3,\ 4) \tag{A.6}$$

This way, it is possible to build a permutation group based on the permutations equivalent to the symmetry operations of the symmetry groups. The permutation group presented in Appendix B was built using the symmetry operations of the I symmetry group on the numbered Deltoidal Hexecontahedron shown in Figure 2.3. The numbering of the faces of the Deltoidal Hexecontahedron followed the numbering used by most $T = 1$ capsids found in *PDB* entries of virus capsids. The permutation group isomorphic to the I permutation group is the Alternating group A_5, with order 60, the same number of symmetry operations on the I symmetry group (which can be seen on Table A.1).

A.2 Burnside's Lemma and its Application

The lemma that gives name to this appendix is the following:

Burnside's Lemma or Cauchy–Frobenius Lemma. *Let G be a group acting on a set S of points. For $g \in G$, let* $\text{fit}(g) = \{\alpha \in S : \alpha \bullet g = \alpha\}$*, being the set of fixed points of S by g. Then:*

$$|G| \cdot |S/G| = \sum_{g \in G} |\text{fix}(g)|$$

or

$$|S/G| = \frac{1}{|G|} \sum_{g \in G} |\text{fix}(g)|$$

Since it is not easy to understand the applications of this lemma in a practical way, and knowing this is not a mathematics dissertation, it will be given an example of the use of this lemma and the algorithmic way used in this work. Taking the example of the square and knowing its rotational symmetry group (C_4 - check the symmetry operations of this group in Table A.1), it can be built the permutation group with cyclical permutations, as we can see in Table A.2.

Definition: Permutation Type. *Let G be a group of permutations on a set S of size n. A permutation $g \in G$ can be described by its* ***type****, defined as the ordered set $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ where λ_i is the number of i-cycles with size i.*

Definition: Cycle Index. *Let G be a group of permutations on a set S of size n. The* ***cycle index*** *is the polynomial with n variables $\nu_1, \nu_2, \ldots, \nu_n$*

$$Z_G(\nu_1, \nu_2, \ldots, \nu_n) = \frac{1}{|G|} \sum_{g \in G} \nu_1^{\lambda_1} \nu_2^{\lambda_2} \cdots \nu_n^{\lambda_n}$$

or

$$Z_G(\nu_1, \nu_2, \ldots, \nu_n) = \frac{1}{|G|} \sum_{g \in G} \prod_{i=1}^{n} \nu_i^{\lambda_i}$$

where the product $\nu_1^{\lambda_1} \nu_2^{\lambda_2} \cdots \nu_n^{\lambda_n}$ is formed for each $g \in G$ from its type $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$.

Table A.2: Cyclical Permutation on the C_4 symmetry group and their Cycle Indexes.

Symmetry Operation	Cyclical Permutation	$\lvert\text{Fix}(g)\rvert$	Permutation Type	Cycle Index Element
E	(1) (2) (3) (4)	c^4	$\{4,0,0,0\}$	v_1^4
C_2	(1, 3) (2, 4)	c^2	$\{0,2,0,0\}$	v_2^2
C_4	(1, 2, 3, 4)	c^1	$\{0,0,0,1\}$	v_4^1
C_4^3	(1, 4, 3, 2)			

$$\begin{aligned} |S/G| &= \frac{1}{|G|}\sum_{g\in G}|\text{fix}(g)| = \\ &= \frac{1}{4}(c^4+c^2+2c) = \frac{1}{4}(2^4+2^2+2\times 2) = 6 \end{aligned} \tag{A.7}$$

$$Z_{C_4}(v_1,v_2,v_3,v_4) = \frac{1}{4}(v_1^4+v_2^2+2v_4) \tag{A.8}$$

Using equation A.8, and considering we will paint this vertices with two colours α and β, it can be substituted v_i for $(\alpha^i+\beta^i)$. Doing so, we will get a polynomial where the coefficients are the number of possible different paintings of the vertices with the combination $\alpha^\eta\beta^{V-\eta}$, $\eta = \{Vertices\ painted\ with\ colour\alpha\}$, $V = |set(Vertices)|$.

$$\begin{aligned} Z_{C_4}(v_1,v_2,v_3,v_4) &= \frac{1}{4}(v_1^4+v_2^2+2v_4) = \\ &= \frac{1}{4}((\alpha+\beta)^4+(\alpha^2+\beta^2)^2+2(\alpha^4+\beta^4)) = \\ &= \frac{1}{4}(4\alpha^4+4\alpha^3\beta+8\alpha^2\beta^2+4\alpha\beta^3+4\beta^4) = \\ &= \alpha^4+\alpha^3\beta+2\alpha^2\beta^2+\alpha\beta^3+\beta^4 \end{aligned} \tag{A.9}$$

This allowed the confirmation that the number of combinations generated using the symmetry group of Appendix B is the same as the number predicted by the lemma for the Deltoidal Hexecontahedron. The cycle index is presented on Equation A.10 and, for the removal of up to 5 proteins, the number of combinations predicted by the lemma and those obtained using the symmetry group were the same.

$$Z_I = \frac{1}{60}(v_1^{60} + 24v_5^{12} + 20v_3^{20} + 15v_2^{30}) \tag{A.10}$$

Number of Combinations Predicted. *Let ρ be the present faces of the Deltoidal Hexecontahedron and let α be the absent ones. Let $v_i = (\rho^i + \alpha^i)$. Then,*

$$Z_I = \frac{1}{60}(v_1^{60} + 24v_5^{12} + 20v_3^{20} + 15v_2^{30}) =$$

$$= \frac{1}{60}\Big((\rho+\alpha)^{60} + 24(\rho^5+\alpha^5)^{12} + 20(\rho^3+\alpha^3)^{20} + 15(\rho^2+\alpha^2)^{30}\Big)$$
$$= \rho^{60} + \rho^{59}\alpha + 37\rho^{58}\alpha^2 + 577\rho^{57}\alpha^3 + 8236\rho^{56}\alpha^4 + 91030\rho^{55}\alpha^5 + \ldots$$

Appendix B

Permutation Group of the *I* Symmetric Group on a Deltoidal Hexecontahedron

E

(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) (28) (29) (30) (31) (32) (33) (34) (35) (36) (37) (38) (39) (40) (41) (42) (43) (44) (45) (46) (47) (48) (49) (50) (51) (52) (53) (54) (55) (56) (57) (58) (59) (60)

$15C_2$

(1, 29) (2, 30) (3, 26) (4, 27) (5, 28) (6, 15) (7, 11) (8, 12) (9, 13) (10, 14) (16, 33) (17, 34) (18, 35) (19, 31) (20, 32) (21, 42) (22, 43) (23, 44) (24, 45) (25, 41) (36, 60) (37, 56) (38, 57) (39, 58) (40, 59) (46, 54) (47, 55) (48, 51) (49, 52) (50, 53)

(1, 47) (2, 48) (3, 49) (4, 50) (5, 46) (6, 13) (7, 14) (8, 15) (9, 11) (10, 12) (16, 55) (17, 51) (18, 52) (19, 53) (20, 54) (21, 40) (22, 36) (23, 37) (24, 38) (25, 39) (26, 59) (27, 60) (28, 56) (29, 57) (30, 58) (31, 43) (32, 44) (33, 45) (34, 41) (35, 42)

(1, 16) (2, 17) (3, 18) (4, 19) (5, 20) (6, 11) (7, 12) (8, 13) (9, 14) (10, 15) (21, 26) (22, 27) (23, 28) (24, 29) (25, 30) (31, 36) (32, 37) (33, 38) (34, 39) (35, 40) (41, 51) (42, 52) (43, 53) (44, 54) (45, 55) (46, 56) (47, 57) (48, 58) (49, 59) (50, 60)

(1, 19) (2, 20) (3, 16) (4, 17) (5, 18) (6, 60) (7, 56) (8, 57) (9, 58) (10, 59) (11, 42) (12, 43) (13,

44) (14, 45) (15, 41) (21, 37) (22, 38) (23, 39) (24, 40) (25, 36) (26, 48) (27, 49) (28, 50) (29, 46) (30, 47) (31, 54) (32, 55) (33, 51) (34, 52) (35, 53)

(1, 11) (2, 12) (3, 13) (4, 14) (5, 15) (6, 16) (7, 17) (8, 18) (9, 19) (10, 20) (21, 36) (22, 37) (23, 38) (24, 39) (25, 40) (26, 31) (27, 32) (28, 33) (29, 34) (30, 35) (41, 46) (42, 47) (43, 48) (44, 49) (45, 50) (51, 56) (52, 57) (53, 58) (54, 59) (55, 60)

(1, 18) (2, 19) (3, 20) (4, 16) (5, 17) (6, 52) (7, 53) (8, 54) (9, 55) (10, 51) (11, 50) (12, 46) (13, 47) (14, 48) (15, 49) (21, 58) (22, 59) (23, 60) (24, 56) (25, 57) (26, 32) (27, 33) (28, 34) (29, 35) (30, 31) (36, 44) (37, 45) (38, 41) (39, 42) (40, 43)

(1, 17) (2, 18) (3, 19) (4, 20) (5, 16) (6, 28) (7, 29) (8, 30) (9, 26) (10, 27) (11, 34) (12, 35) (13, 31) (14, 32) (15, 33) (21, 55) (22, 51) (23, 52) (24, 53) (25, 54) (36, 47) (37, 48) (38, 49) (39, 50) (40, 46) (41, 59) (42, 60) (43, 56) (44, 57) (45, 58)

(1, 6) (2, 7) (3, 8) (4, 9) (5, 10) (11, 16) (12, 17) (13, 18) (14, 19) (15, 20) (21, 31) (22, 32) (23, 33) (24, 34) (25, 35) (26, 36) (27, 37) (28, 38) (29, 39) (30, 40) (41, 56) (42, 57) (43, 58) (44, 59) (45, 60) (46, 51) (47, 52) (48, 53) (49, 54) (50, 55)

(1, 35) (2, 31) (3, 32) (4, 33) (5, 34) (6, 49) (7, 50) (8, 46) (9, 47) (10, 48) (11, 53) (12, 54) (13, 55) (14, 51) (15, 52) (16, 27) (17, 28) (18, 29) (19, 30) (20, 26) (21, 39) (22, 40) (23, 36) (24, 37) (25, 38) (41, 57) (42, 58) (43, 59) (44, 60) (45, 56)

(1, 45) (2, 41) (3, 42) (4, 43) (5, 44) (6, 14) (7, 15) (8, 11) (9, 12) (10, 13) (16, 57) (17, 58) (18, 59) (19, 60) (20, 56) (21, 49) (22, 50) (23, 46) (24, 47) (25, 48) (26, 35) (27, 31) (28, 32) (29, 33) (30, 34) (36, 53) (37, 54) (38, 55) (39, 51) (40, 52)

(1, 54) (2, 55) (3, 51) (4, 52) (5, 53) (6, 25) (7, 21) (8, 22) (9, 23) (10, 24) (11, 37) (12, 38) (13, 39) (14, 40) (15, 36) (16, 48) (17, 49) (18, 50) (19, 46) (20, 47) (26, 33) (27, 34) (28, 35) (29, 31) (30, 32) (41, 60) (42, 56) (43, 57) (44, 58) (45, 59)

(1, 38) (2, 39) (3, 40) (4, 36) (5, 37) (6, 12) (7, 13) (8, 14) (9, 15) (10, 11) (16, 24) (17, 25) (18, 21) (19, 22) (20, 23) (26, 52) (27, 53) (28, 54) (29, 55) (30, 51) (31, 50) (32, 46) (33, 47) (34, 48) (35, 49) (41, 58) (42, 59) (43, 60) (44, 56) (45, 57)

(1, 22) (2, 23) (3, 24) (4, 25) (5, 21) (6, 43) (7, 44) (8, 45) (9, 41) (10, 42) (11, 59) (12, 60) (13, 56) (14, 57) (15, 58) (16, 40) (17, 36) (18, 37) (19, 38) (20, 39) (26, 34) (27, 35) (28, 31) (29, 32) (30, 33) (46, 55) (47, 51) (48, 52) (49, 53) (50, 54)

(1, 20) (2, 16) (3, 17) (4, 18) (5, 19) (6, 39) (7, 40) (8, 36) (9, 37) (10, 38) (11, 23) (12, 24) (13, 25) (14, 21) (15, 22) (26, 45) (27, 41) (28, 42) (29, 43) (30, 44) (31, 57) (32, 58) (33, 59) (34, 60) (35,

56) (46, 53) (47, 54) (48, 55) (49, 51) (50, 52)

(1, 58) (2, 59) (3, 60) (4, 56) (5, 57) (6, 32) (7, 33) (8, 34) (9, 35) (10, 31) (11, 30) (12, 26) (13, 27) (14, 28) (15, 29) (16, 44) (17, 45) (18, 41) (19, 42) (20, 43) (21, 38) (22, 39) (23, 40) (24, 36) (25, 37) (46, 52) (47, 53) (48, 54) (49, 55) (50, 51)

$20C_3$

(1, 60, 14) (2, 56, 15) (3, 57, 11) (4, 58, 12) (5, 59, 13) (6, 19, 45) (7, 20, 41) (8, 16, 42) (9, 17, 43) (10, 18, 44) (21, 54, 27) (22, 55, 28) (23, 51, 29) (24, 52, 30) (25, 53, 26) (31, 37, 49) (32, 38, 50) (33, 39, 46) (34, 40, 47) (35, 36, 48)

(1, 51, 36) (2, 52, 37) (3, 53, 38) (4, 54, 39) (5, 55, 40) (6, 56, 31) (7, 57, 32) (8, 58, 33) (9, 59, 34) (10, 60, 35) (11, 41, 26) (12, 42, 27) (13, 43, 28) (14, 44, 29) (15, 45, 30) (16, 46, 21) (17, 47, 22) (18, 48, 23) (19, 49, 24) (20, 50, 25)

(1, 34, 7) (2, 35, 8) (3, 31, 9) (4, 32, 10) (5, 33, 6) (11, 17, 29) (12, 18, 30) (13, 19, 26) (14, 20, 27) (15, 16, 28) (21, 47, 60) (22, 48, 56) (23, 49, 57) (24, 50, 58) (25, 46, 59) (36, 55, 42) (37, 51, 43) (38, 52, 44) (39, 53, 45) (40, 54, 41)

(1, 10, 23) (2, 6, 24) (3, 7, 25) (4, 8, 21) (5, 9, 22) (11, 38, 20) (12, 39, 16) (13, 40, 17) (14, 36, 18) (15, 37, 19) (26, 50, 57) (27, 46, 58) (28, 47, 59) (29, 48, 60) (30, 49, 56) (31, 52, 45) (32, 53, 41) (33, 54, 42) (34, 55, 43) (35, 51, 44)

(1, 52, 13) (2, 53, 14) (3, 54, 15) (4, 55, 11) (5, 51, 12) (6, 18, 47) (7, 19, 48) (8, 20, 49) (9, 16, 50) (10, 17, 46) (21, 30, 43) (22, 26, 44) (23, 27, 45) (24, 28, 41) (25, 29, 42) (31, 58, 40) (32, 59, 36) (33, 60, 37) (34, 56, 38) (35, 57, 39)

(1, 21, 41) (2, 22, 42) (3, 23, 43) (4, 24, 44) (5, 25, 45) (6, 26, 46) (7, 27, 47) (8, 28, 48) (9, 29, 49) (10, 30, 50) (11, 31, 51) (12, 32, 52) (13, 33, 53) (14, 34, 54) (15, 35, 55) (16, 36, 56) (17, 37, 57) (18, 38, 58) (19, 39, 59) (20, 40, 60)

(1, 46, 31) (2, 47, 32) (3, 48, 33) (4, 49, 34) (5, 50, 35) (6, 41, 36) (7, 42, 37) (8, 43, 38) (9, 44, 39) (10, 45, 40) (11, 56, 21) (12, 57, 22) (13, 58, 23) (14, 59, 24) (15, 60, 25) (16, 51, 26) (17, 52, 27) (18, 53, 28) (19, 54, 29) (20, 55, 30)

(1, 56, 26) (2, 57, 27) (3, 58, 28) (4, 59, 29) (5, 60, 30) (6, 51, 21) (7, 52, 22) (8, 53, 23) (9, 54, 24) (10, 55, 25) (11, 46, 36) (12, 47, 37) (13, 48, 38) (14, 49, 39) (15, 50, 40) (16, 41, 31) (17, 42, 32) (18, 43, 33) (19, 44, 34) (20, 45, 35)

(1, 27, 40) (2, 28, 36) (3, 29, 37) (4, 30, 38) (5, 26, 39) (6, 53, 59) (7, 54, 60) (8, 55, 56) (9, 51,

57) (10, 52, 58) (11, 49, 43) (12, 50, 44) (13, 46, 45) (14, 47, 41) (15, 48, 42) (16, 35, 22) (17, 31, 23) (18, 32, 24) (19, 33, 25) (20, 34, 21)

(1, 44, 48) (2, 45, 49) (3, 41, 50) (4, 42, 46) (5, 43, 47) (6, 30, 37) (7, 26, 38) (8, 27, 39) (9, 28, 40) (10, 29, 36) (11, 32, 25) (12, 33, 21) (13, 34, 22) (14, 35, 23) (15, 31, 24) (16, 58, 54) (17, 59, 55) (18, 60, 51) (19, 56, 52) (20, 57, 53)

(1, 14, 60) (2, 15, 56) (3, 11, 57) (4, 12, 58) (5, 13, 59) (6, 45, 19) (7, 41, 20) (8, 42, 16) (9, 43, 17) (10, 44, 18) (21, 27, 54) (22, 28, 55) (23, 29, 51) (24, 30, 52) (25, 26, 53) (31, 49, 37) (32, 50, 38) (33, 46, 39) (34, 47, 40) (35, 48, 36)

(1, 40, 27) (2, 36, 28) (3, 37, 29) (4, 38, 30) (5, 39, 26) (6, 59, 53) (7, 60, 54) (8, 56, 55) (9, 57, 51) (10, 58, 52) (11, 43, 49) (12, 44, 50) (13, 45, 46) (14, 41, 47) (15, 42, 48) (16, 22, 35) (17, 23, 31) (18, 24, 32) (19, 25, 33) (20, 21, 34)

(1, 13, 52) (2, 14, 53) (3, 15, 54) (4, 11, 55) (5, 12, 51) (6, 47, 18) (7, 48, 19) (8, 49, 20) (9, 50, 16) (10, 46, 17) (21, 43, 30) (22, 44, 26) (23, 45, 27) (24, 41, 28) (25, 42, 29) (31, 40, 58) (32, 36, 59) (33, 37, 60) (34, 38, 56) (35, 39, 57)

(1, 23, 10) (2, 24, 6) (3, 25, 7) (4, 21, 8) (5, 22, 9) (11, 20, 38) (12, 16, 39) (13, 17, 40) (14, 18, 36) (15, 19, 37) (26, 57, 50) (27, 58, 46) (28, 59, 47) (29, 60, 48) (30, 56, 49) (31, 45, 52) (32, 41, 53) (33, 42, 54) (34, 43, 55) (35, 44, 51)

(1, 26, 56) (2, 27, 57) (3, 28, 58) (4, 29, 59) (5, 30, 60) (6, 21, 51) (7, 22, 52) (8, 23, 53) (9, 24, 54) (10, 25, 55) (11, 36, 46) (12, 37, 47) (13, 38, 48) (14, 39, 49) (15, 40, 50) (16, 31, 41) (17, 32, 42) (18, 33, 43) (19, 34, 44) (20, 35, 45)

(1, 7, 34) (2, 8, 35) (3, 9, 31) (4, 10, 32) (5, 6, 33) (11, 29, 17) (12, 30, 18) (13, 26, 19) (14, 27, 20) (15, 28, 16) (21, 60, 47) (22, 56, 48) (23, 57, 49) (24, 58, 50) (25, 59, 46) (36, 42, 55) (37, 43, 51) (38, 44, 52) (39, 45, 53) (40, 41, 54)

(1, 36, 51) (2, 37, 52) (3, 38, 53) (4, 39, 54) (5, 40, 55) (6, 31, 56) (7, 32, 57) (8, 33, 58) (9, 34, 59) (10, 35, 60) (11, 26, 41) (12, 27, 42) (13, 28, 43) (14, 29, 44) (15, 30, 45) (16, 21, 46) (17, 22, 47) (18, 23, 48) (19, 24, 49) (20, 25, 50)

(1, 41, 21) (2, 42, 22) (3, 43, 23) (4, 44, 24) (5, 45, 25) (6, 46, 26) (7, 47, 27) (8, 48, 28) (9, 49, 29) (10, 50, 30) (11, 51, 31) (12, 52, 32) (13, 53, 33) (14, 54, 34) (15, 55, 35) (16, 56, 36) (17, 57, 37) (18, 58, 38) (19, 59, 39) (20, 60, 40)

(1, 31, 46) (2, 32, 47) (3, 33, 48) (4, 34, 49) (5, 35, 50) (6, 36, 41) (7, 37, 42) (8, 38, 43) (9, 39, 44) (10, 40, 45) (11, 21, 56) (12, 22, 57) (13, 23, 58) (14, 24, 59) (15, 25, 60) (16, 26, 51) (17, 27, 52)

(18, 28, 53) (19, 29, 54) (20, 30, 55)

(1, 48, 44) (2, 49, 45) (3, 50, 41) (4, 46, 42) (5, 47, 43) (6, 37, 30) (7, 38, 26) (8, 39, 27) (9, 40, 28) (10, 36, 29) (11, 25, 32) (12, 21, 33) (13, 22, 34) (14, 23, 35) (15, 24, 31) (16, 54, 58) (17, 55, 59) (18, 51, 60) (19, 52, 56) (20, 53, 57)

$$12\mathrm{C}_5 + 12C_5^2$$

(1, 43, 15, 39, 32) (2, 44, 11, 40, 33) (3, 45, 12, 36, 34) (4, 41, 13, 37, 35) (5, 42, 14, 38, 31) (6, 22, 29, 20, 58) (7, 23, 30, 16, 59) (8, 24, 26, 17, 60) (9, 25, 27, 18, 56) (10, 21, 28, 19, 57) (46, 47, 48, 49, 50) (51, 55, 54, 53, 52)

(1, 25, 28, 12, 49) (2, 21, 29, 13, 50) (3, 22, 30, 14, 46) (4, 23, 26, 15, 47) (5, 24, 27, 11, 48) (6, 54, 17, 38, 35) (7, 55, 18, 39, 31) (8, 51, 19, 40, 32) (9, 52, 20, 36, 33) (10, 53, 16, 37, 34) (41, 42, 43, 44, 45) (56, 60, 59, 58, 57)

(1, 9, 53, 30, 42) (2, 10, 54, 26, 43) (3, 6, 55, 27, 44) (4, 7, 51, 28, 45) (5, 8, 52, 29, 41) (11, 47, 35, 58, 19) (12, 48, 31, 59, 20) (13, 49, 32, 60, 16) (14, 50, 33, 56, 17) (15, 46, 34, 57, 18) (21, 22, 23, 24, 25) (36, 40, 39, 38, 37)

(1, 2, 3, 4, 5) (6, 23, 42, 50, 34) (7, 24, 43, 46, 35) (8, 25, 44, 47, 31) (9, 21, 45, 48, 32) (10, 22, 41, 49, 33) (11, 39, 60, 52, 28) (12, 40, 56, 53, 29) (13, 36, 57, 54, 30) (14, 37, 58, 55, 26) (15, 38, 59, 51, 27) (16, 20, 19, 18, 17)

(1, 33, 57, 55, 24) (2, 34, 58, 51, 25) (3, 35, 59, 52, 21) (4, 31, 60, 53, 22) (5, 32, 56, 54, 23) (6, 7, 8, 9, 10) (11, 15, 14, 13, 12) (16, 29, 45, 47, 38) (17, 30, 41, 48, 39) (18, 26, 42, 49, 40) (19, 27, 43, 50, 36) (20, 28, 44, 46, 37)

(1, 50, 37, 59, 8) (2, 46, 38, 60, 9) (3, 47, 39, 56, 10) (4, 48, 40, 57, 6) (5, 49, 36, 58, 7) (11, 18, 54, 22, 45) (12, 19, 55, 23, 41) (13, 20, 51, 24, 42) (14, 16, 52, 25, 43) (15, 17, 53, 21, 44) (26, 30, 29, 28, 27) (31, 32, 33, 34, 35)

(1, 39, 43, 32, 15) (2, 40, 44, 33, 11) (3, 36, 45, 34, 12) (4, 37, 41, 35, 13) (5, 38, 42, 31, 14) (6, 20, 22, 58, 29) (7, 16, 23, 59, 30) (8, 17, 24, 60, 26) (9, 18, 25, 56, 27) (10, 19, 21, 57, 28) (46, 49, 47, 50, 48) (51, 53, 55, 52, 54)

(1, 3, 5, 2, 4) (6, 42, 34, 23, 50) (7, 43, 35, 24, 46) (8, 44, 31, 25, 47) (9, 45, 32, 21, 48) (10, 41, 33, 22, 49) (11, 60, 28, 39, 52) (12, 56, 29, 40, 53) (13, 57, 30, 36, 54) (14, 58, 26, 37, 55) (15, 59, 27, 38, 51) (16, 19, 17, 20, 18)

(1, 59, 50, 8, 37) (2, 60, 46, 9, 38) (3, 56, 47, 10, 39) (4, 57, 48, 6, 40) (5, 58, 49, 7, 36) (11, 22,

18, 45, 54) (12, 23, 19, 41, 55) (13, 24, 20, 42, 51) (14, 25, 16, 43, 52) (15, 21, 17, 44, 53) (26, 28, 30, 27, 29) (31, 34, 32, 35, 33)

(1, 53, 42, 9, 30) (2, 54, 43, 10, 26) (3, 55, 44, 6, 27) (4, 51, 45, 7, 28) (5, 52, 41, 8, 29) (11, 35, 19, 47, 58) (12, 31, 20, 48, 59) (13, 32, 16, 49, 60) (14, 33, 17, 50, 56) (15, 34, 18, 46, 57) (21, 23, 25, 22, 24) (36, 39, 37, 40, 38)

(1, 28, 49, 25, 12) (2, 29, 50, 21, 13) (3, 30, 46, 22, 14) (4, 26, 47, 23, 15) (5, 27, 48, 24, 11) (6, 17, 35, 54, 38) (7, 18, 31, 55, 39) (8, 19, 32, 51, 40) (9, 20, 33, 52, 36) (10, 16, 34, 53, 37) (41, 43, 45, 42, 44) (56, 59, 57, 60, 58)

(1, 4, 2, 5, 3) (6, 50, 23, 34, 42) (7, 46, 24, 35, 43) (8, 47, 25, 31, 44) (9, 48, 21, 32, 45) (10, 49, 22, 33, 41) (11, 52, 39, 28, 60) (12, 53, 40, 29, 56) (13, 54, 36, 30, 57) (14, 55, 37, 26, 58) (15, 51, 38, 27, 59) (16, 18, 20, 17, 19)

(1, 42, 30, 53, 9) (2, 43, 26, 54, 10) (3, 44, 27, 55, 6) (4, 45, 28, 51, 7) (5, 41, 29, 52, 8) (11, 19, 58, 35, 47) (12, 20, 59, 31, 48) (13, 16, 60, 32, 49) (14, 17, 56, 33, 50) (15, 18, 57, 34, 46) (21, 25, 24, 23, 22) (36, 37, 38, 39, 40)

(1, 30, 9, 42, 53) (2, 26, 10, 43, 54) (3, 27, 6, 44, 55) (4, 28, 7, 45, 51) (5, 29, 8, 41, 52) (11, 58, 47, 19, 35) (12, 59, 48, 20, 31) (13, 60, 49, 16, 32) (14, 56, 50, 17, 33) (15, 57, 46, 18, 34) (21, 24, 22, 25, 23) (36, 38, 40, 37, 39)

(1, 57, 24, 33, 55) (2, 58, 25, 34, 51) (3, 59, 21, 35, 52) (4, 60, 22, 31, 53) (5, 56, 23, 32, 54) (6, 8, 10, 7, 9) (11, 14, 12, 15, 13) (16, 45, 38, 29, 47) (17, 41, 39, 30, 48) (18, 42, 40, 26, 49) (19, 43, 36, 27, 50) (20, 44, 37, 28, 46)

(1, 55, 33, 24, 57) (2, 51, 34, 25, 58) (3, 52, 35, 21, 59) (4, 53, 31, 22, 60) (5, 54, 32, 23, 56) (6, 9, 7, 10, 8) (11, 13, 15, 12, 14) (16, 47, 29, 38, 45) (17, 48, 30, 39, 41) (18, 49, 26, 40, 42) (19, 50, 27, 36, 43) (20, 46, 28, 37, 44)

(1, 8, 59, 37, 50) (2, 9, 60, 38, 46) (3, 10, 56, 39, 47) (4, 6, 57, 40, 48) (5, 7, 58, 36, 49) (11, 45, 22, 54, 18) (12, 41, 23, 55, 19) (13, 42, 24, 51, 20) (14, 43, 25, 52, 16) (15, 44, 21, 53, 17) (26, 27, 28, 29, 30) (31, 35, 34, 33, 32)

(1, 49, 12, 28, 25) (2, 50, 13, 29, 21) (3, 46, 14, 30, 22) (4, 47, 15, 26, 23) (5, 48, 11, 27, 24) (6, 35, 38, 17, 54) (7, 31, 39, 18, 55) (8, 32, 40, 19, 51) (9, 33, 36, 20, 52) (10, 34, 37, 16, 53) (41, 45, 44, 43, 42) (56, 57, 58, 59, 60)

(1, 12, 25, 49, 28) (2, 13, 21, 50, 29) (3, 14, 22, 46, 30) (4, 15, 23, 47, 26) (5, 11, 24, 48, 27) (6, 38, 54, 35, 17) (7, 39, 55, 31, 18) (8, 40, 51, 32, 19) (9, 36, 52, 33, 20) (10, 37, 53, 34, 16) (41, 44, 42,

45, 43) (56, 58, 60, 57, 59)

(1, 32, 39, 15, 43) (2, 33, 40, 11, 44) (3, 34, 36, 12, 45) (4, 35, 37, 13, 41) (5, 31, 38, 14, 42) (6, 58, 20, 29, 22) (7, 59, 16, 30, 23) (8, 60, 17, 26, 24) (9, 56, 18, 27, 25) (10, 57, 19, 28, 21) (46, 50, 49, 48, 47) (51, 52, 53, 54, 55)

(1, 24, 55, 57, 33) (2, 25, 51, 58, 34) (3, 21, 52, 59, 35) (4, 22, 53, 60, 31) (5, 23, 54, 56, 32) (6, 10, 9, 8, 7) (11, 12, 13, 14, 15) (16, 38, 47, 45, 29) (17, 39, 48, 41, 30) (18, 40, 49, 42, 26) (19, 36, 50, 43, 27) (20, 37, 46, 44, 28)

(1, 5, 4, 3, 2) (6, 34, 50, 42, 23) (7, 35, 46, 43, 24) (8, 31, 47, 44, 25) (9, 32, 48, 45, 21) (10, 33, 49, 41, 22) (11, 28, 52, 60, 39) (12, 29, 53, 56, 40) (13, 30, 54, 57, 36) (14, 26, 55, 58, 37) (15, 27, 51, 59, 38) (16, 17, 18, 19, 20)

(1, 37, 8, 50, 59) (2, 38, 9, 46, 60) (3, 39, 10, 47, 56) (4, 40, 6, 48, 57) (5, 36, 7, 49, 58) (11, 54, 45, 18, 22) (12, 55, 41, 19, 23) (13, 51, 42, 20, 24) (14, 52, 43, 16, 25) (15, 53, 44, 17, 21) (26, 29, 27, 30, 28) (31, 33, 35, 32, 34)

(1, 15, 32, 43, 39) (2, 11, 33, 44, 40) (3, 12, 34, 45, 36) (4, 13, 35, 41, 37) (5, 14, 31, 42, 38) (6, 29, 58, 22, 20) (7, 30, 59, 23, 16) (8, 26, 60, 24, 17) (9, 27, 56, 25, 18) (10, 28, 57, 21, 19) (46, 48, 50, 47, 49) (51, 54, 52, 55, 53)

Appendix C

Supplementary Results using Heuristic I

Table C.1: Results when five proteins are removed, for Adeno-Associated Virus Capsid structures, using the heuristic I.

	PDBID	Energy Minimum	Second Energy Minimum
Adeno-Associated Virus	3j4p	$\{1,2,10,22,23\}$	$\{1,2,10,23,42\}$
	3ntt		
	3ux1		
	4iov		
	3j1q		
	3ra2		$\{1,2,6,10,23\}$
	3ra8		
	2g8g	$\{1,6,25,26,53\}$ $\{1,6,13,45,46\}$	
	1lp3	$\{1,2,6,10,23\}$	$\{1,2,10,22,23\}$
	2qa0		
	3ra4		
	3ra9		
	3raa		
	4rso		
	5egc		

Table C.2: Results when five proteins are removed, for Bovine, Human and Porcine Parvoviruses Capsid structures, using the heuristic I.

	PDBID	Energy Minimum	Second Energy Minimum
Bovine Parvovirus	4qc8	$\{1,2,10,22,23\}$	$\{1,2,10,23,42\}$
Human Parvovirus	1s58		
Porcine Parvovirus	1k3v		$\{1,10,11,20,23\}$

Table C.3: Results when four ($N = 4$) and five ($N = 5$) proteins are removed, for *Bombyx mori*, *Galleria mellonella* and *Penaeus stylirostris* Densoviruses Capsid structures, using the heuristic I.

		Energy Minimum	Second Energy Minimum
	PDBID	$N = 4$	
Bombyx mori Densovirus	3p0s	$\{1,2,22,23\}$	$\{1,2,10,23\}$
Galleria mellonella Densovirus	1dnv		
Penaeus stylirostris Densovirus	3n7x		
		$N = 5$	
Bombyx mori Densovirus	3p0s	$\{1,2,6,10,23\}$	$\{1,2,10,22,23\}$
Galleria mellonella Densovirus	1dnv	$\{1,2,10,22,23\}$	$\{1,2,6,10,23\}$
Penaeus stylirostris Densovirus	3n7x		$\{1,2,3,41,42\}$

Table C.4: Results when five proteins are removed, for Hepatitis E Virus Capsid structures, using the heuristic I.

	PDBID	Energy Minimum	Second Energy Minimum
Hepatitis E Virus	2ztn	$\{1,2,6,10,23\}$	$\{1,6,25,26,53\}$ $\{1,6,13,45,46\}$
	2zzq		$\{1,6,14,29,45\}$
	3hag	$\{1,2,10,22,23\}$	$\{1,2,6,10,23\}$

Table C.5: Results when five proteins are removed, for Canine and Feline Panleukopenia Virus Capsid structures, using the heuristic I.

	PDBID	Energy Minimum	Second Energy Minimum
Canine & Feline Panleukopenia Virus	1c8f	$\{1,2,10,22,23\}$	$\{1,2,10,23,42\}$
	1c8d		
	1c8e		
	1c8g		
	1fpv		
	1p5y		
	1c8h		$\{1,2,10,22,42\}$
	1ijs		
	1p5w		
	2cas		
	4dpv		

Table C.6: Results when five proteins are removed, for Rodent Protoparvovirus Capsid structures, using the heuristic I.

	PDBID	Energy Minimum	Second Energy Minimum
Rodent Protoparvovirus	1mvm	$\{1,2,10,22,23\}$	$\{1,3,10,41,50\}$ $\{1,10,11,20,23\}$
	1z1c		$\{1,2,6,10,23\}$
	1z14		$\{1,2,10,23,42\}$
	2xgk		$\{1,3,10,41,50\}$ $\{1,2,10,23,42\}$ $\{1,10,11,20,23\}$
	4g0r		$\{1,3,10,41,50\}$ $\{1,10,11,20,23\}$
	4gbt		$\{1,2,3,41,50\}$

Table C.7: Results when five proteins are removed, for Satellite Tobacco Mosaic Virus Capsid structures, using the heuristic I.

	PDBID	Energy Minimum	Second Energy Minimum
Satellite Tobacco Mosaic Virus	1a34	$\{1,2,3,23,42\}$ $\{1,2,6,10,23\}$	$\{1,2,6,22,23\}$
	2buk	$\{1,2,6,10,24\}$	$\{1,2,6,11,16\}$ $\{1,2,6,31,36\}$ $\{1,2,6,21,26\}$
	4bcu	$\{1,2,3,23,42\}$	$\{1,2,6,10,24\}$
	4oq8	$\{1,2,6,10,23\}$	$\{1,2,3,23,42\}$

Table C.8: Results when five proteins are removed, for Avian Birnavirus Capsid structures, using the heuristic I.

	PDBID	Energy Minimum	Second Energy Minimum
Avian Birnavirus	1cwd	$\{1,2,10,22,23\}$	$\{1,2,10,23,42\}$ $\{1,8,23,55,56\}$ $\{1,10,13,23,45\}$ $\{1,10,13,23,46\}$ $\{1,10,15,23,28\}$

Table C.9: Results when five proteins are removed, for Porcine Circovirus Capsid structures, using the heuristic I.

	PDBID	Energy Minimum	Second Energy Minimum
Porcine Circovirus	3jci	$\{1,2,10,22,23\}$	$\{1,2,10,23,42\}$
	3r0r		

Table C.10: Results when five proteins are removed, for Human Adenovirus Pt-Dd Capsid structures, using the heuristic I.

	PDBID	Energy Minimum	Second Energy Minimum
Human Adenovirus Pt-Dd	1x9t	$\{1,2,3,4,5\}$	$\{1,2,3,23,24\}$
	4aqq		$\{1,2,7,8,34\}$
	4ar2		$\{1,2,3,34,35\}$

Appendix D

Supplementary Results using Heuristic II

Table D.1: Results when five proteins are removed, for Adeno-Associated Virus Capsid structures, using the heuristic II.

	PDBID	Energy Minimum	Second Energy Minimum
Adeno-Associated Virus	3j4p	$\{1,2,10,22,23\}$	$\{1,2,10,23,42\}$
	3ntt		
	3ux1		
	4iov		
	3j1q		
	3ra2		
	3ra8		
	2g8g		
	1lp3		$\{1,2,6,10,23\}$
	2qa0		$\{1,2,10,23,42\}$
	3ra4		
	3ra9		
	3raa		
	4rso		$\{1,2,6,10,23\}$
	5egc		$\{1,2,10,23,42\}$ $\{1,2,6,10,23\}$

Table D.2: Results when five proteins are removed, for Bovine, Human and Porcine Parvoviruses Capsid structures, using the heuristic II.

	PDBID	Energy Minimum	Second Energy Minimum
Bovine Parvovirus	4qc8	$\{1,2,10,22,23\}$	$\{1,2,10,23,42\}$
Human Parvovirus	1s58		$\{1,2,6,10,23\}$
Porcine Parvovirus	1k3v		$\{1,2,10,23,42\}$

Table D.3: Results when four ($N = 4$) and five ($N = 5$) proteins are removed, for *Bombyx mori*, *Galleria mellonella* and *Penaeus stylirostris* Densoviruses Capsid structures, using the heuristic II.

		Energy Minimum	Second Energy Minimum
	PDBID	$N = 4$	
Bombyx mori Densovirus	3p0s	$\{1,2,22,23\}$	$\{1,2,10,23\}$
Galleria mellonella Densovirus	1dnv		
Penaeus stylirostris Densovirus	3n7x		
		$N = 5$	
Bombyx mori Densovirus	3p0s	$\{1,2,10,22,23\}$	$\{1,2,6,10,23\}$
Galleria mellonella Densovirus	1dnv		
Penaeus stylirostris Densovirus	3n7x	$\{1,2,6,10,23\}$	$\{1,2,3,23,42\}$

Table D.4: Results when five proteins are removed, for Hepatitis E Virus Capsid structures, using the heuristic II.

	PDBID	Energy Minimum	Second Energy Minimum
Hepatitis E Virus	2ztn	$\{1,2,6,10,23\}$	$\{1,2,6,22,23\}$
	2zzq		
	3hag		$\{1,2,3,23,42\}$

Table D.5: Results when five proteins are removed, for Canine and Feline Panleukopenia Virus Capsid structures, using the heuristic II.

	PDBID	Energy Minimum	Second Energy Minimum
Canine & Feline Panleukopenia Virus	1c8f	$\{1,2,10,22,23\}$	$\{1,10,14,23,44\}$ $\{1,10,23,29,44\}$
	1c8d		$\{1,2,10,23,42\}$
	1c8e		
	1c8g		
	1fpv		
	1p5y		
	1c8h		
	1ijs		
	1p5w		
	2cas		
	4dpv		

Table D.6: Results when five proteins are removed, for Rodent Protoparvovirus Capsid structures, using the heuristic II.

	PDBID	Energy Minimum	Second Energy Minimum
Rodent Protoparvovirus	1mvm	$\{1,2,10,22,23\}$	$\{1,2,10,23,42\}$
	1z1c		
	1z14		
	2xgk		
	4g0r		
	4gbt		

Table D.7: Results when five proteins are removed, for Satellite Tobacco Mosaic Virus Capsid structures, using the heuristic II.

	PDBID	Energy Minimum	Second Energy Minimum
Satellite Tobacco Mosaic Virus	1a34	$\{1,2,3,23,42\}$ $\{1,2,6,10,23\}$	$\{1,2,6,22,23\}$
	2buk	$\{1,2,6,10,24\}$	$\{1,2,6,7,24\}$
	4bcu		
	4oq8	$\{1,2,6,10,23\}$	$\{1,2,3,23,42\}$

Table D.8: Results when five proteins are removed, for Avian Birnavirus Capsid structures, using the heuristic II.

	PDBID	Energy Minimum	Second Energy Minimum
Avian Birnavirus	1cwd	$\{1,2,10,22,23\}$	$\{1,2,10,23,42\}$

Table D.9: Results when five proteins are removed, for Porcine Circovirus Capsid structures, using the heuristic II.

	PDBID	Energy Minimum	Second Energy Minimum
Porcine Circovirus	3jci	$\{1,2,10,22,23\}$	$\{1,2,10,23,42\}$
	3r0r		

Table D.10: Results when five proteins are removed, for Human Adenovirus Pt-Dd Capsid structures, using the heuristic II.

	PDBID	Energy Minimum	Second Energy Minimum
Human Adenovirus Pt-Dd	1x9t	$\{1,2,3,4,5\}$	$\{1,2,3,23,24\}$
	4aqq		$\{1,2,3,6,7\}$
	4ar2		$\{1,2,3,23,24\}$

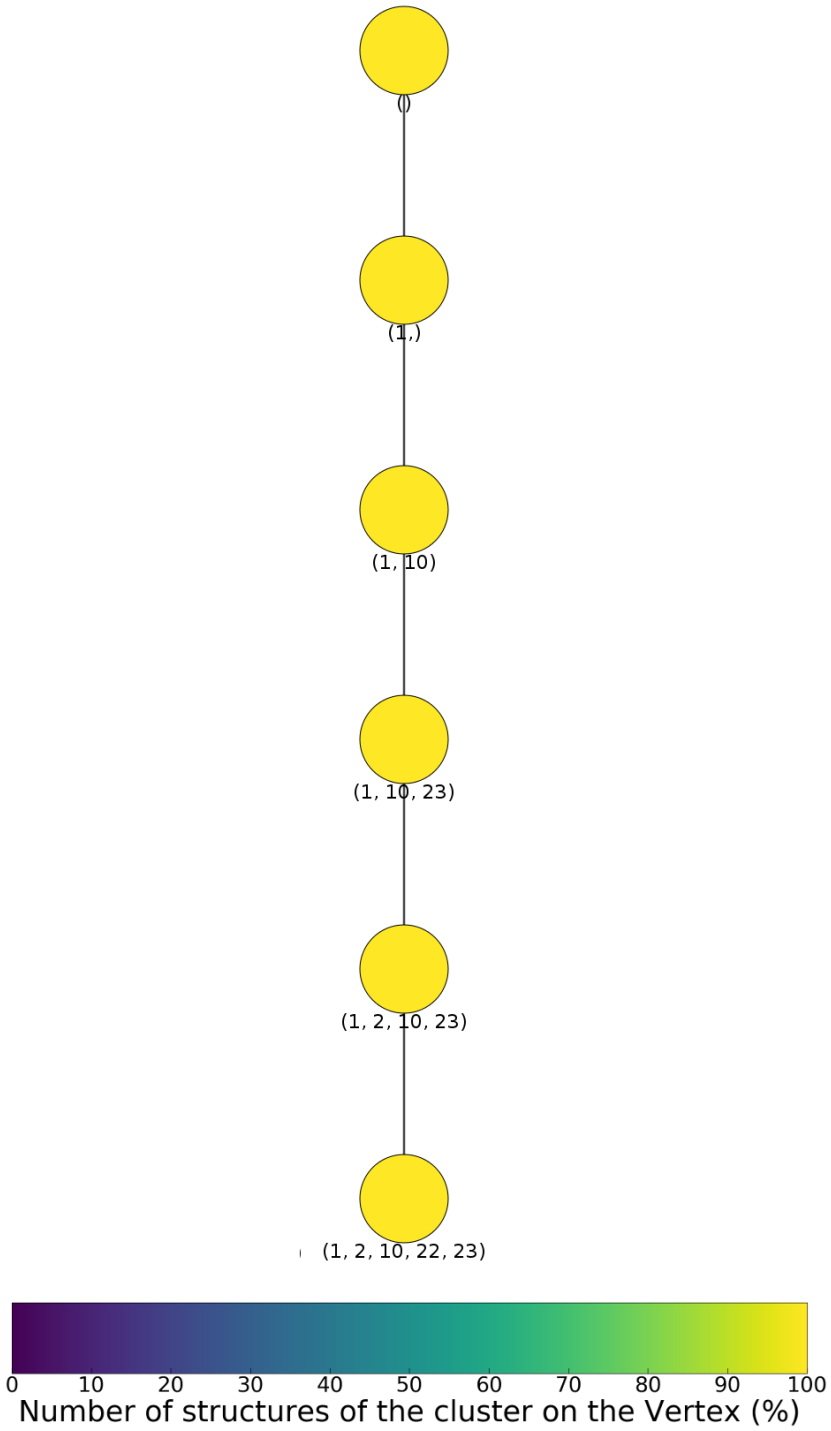

Figure D.1: Disassembly path with the energy minimum protein combination obtained for each PDB structure of the Adeno-Associated Virus, Avian Birnavirus, Bovine Parvovirus, Human Parvovirus, Porcine Parvovirus and Rodent Protoparvovirus groups, in each step of the removal of one protein, with heuristic II.

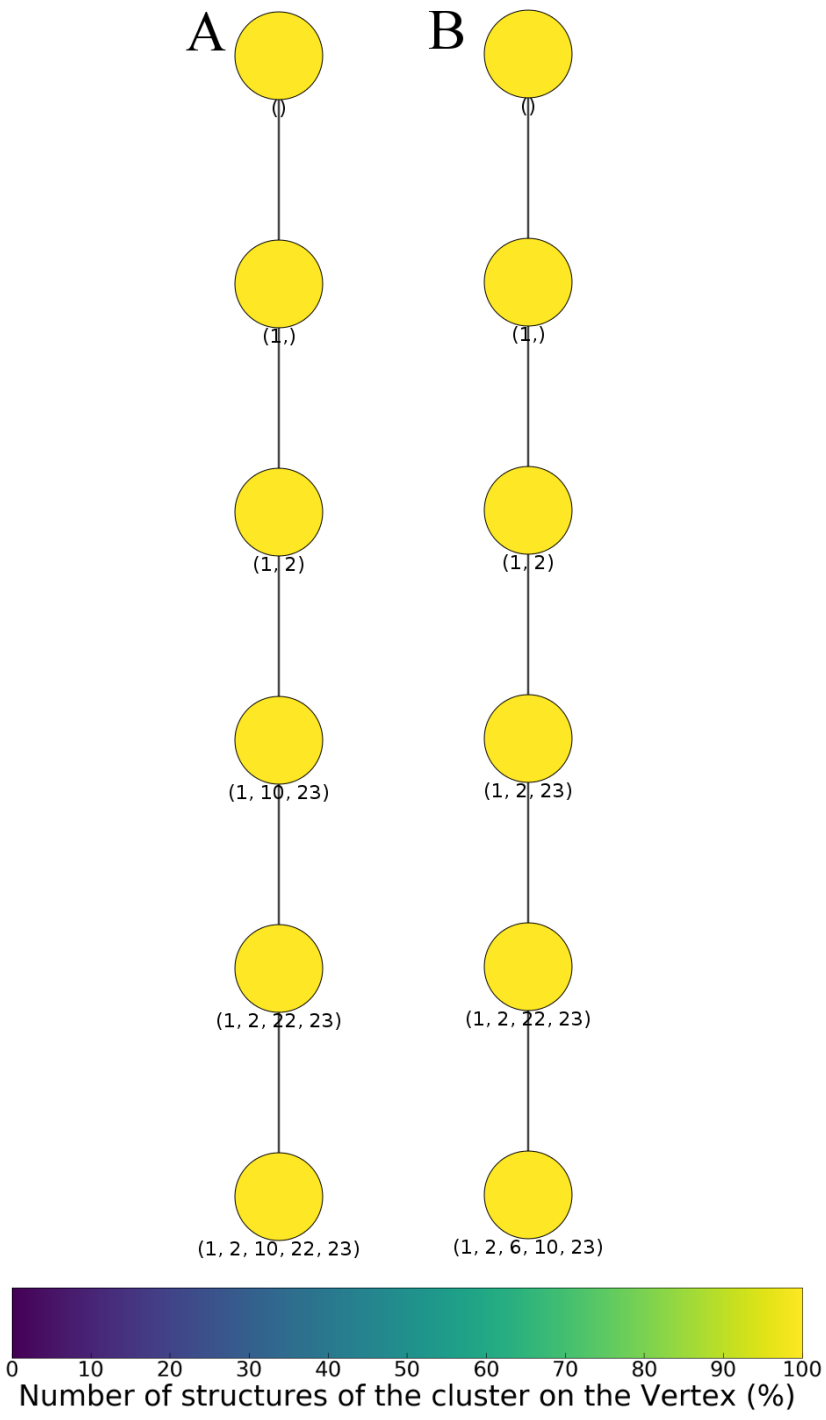

Figure D.2: Disassembly path with the energy minimum protein combination obtained for each PDB structure of the *Bombyx mori* and *Galleria mellonella* Densoviruses groups (A), and *Penaeus stylirostris* Densovirus groups (B), in each step of the removal of one protein, with heuristic II.

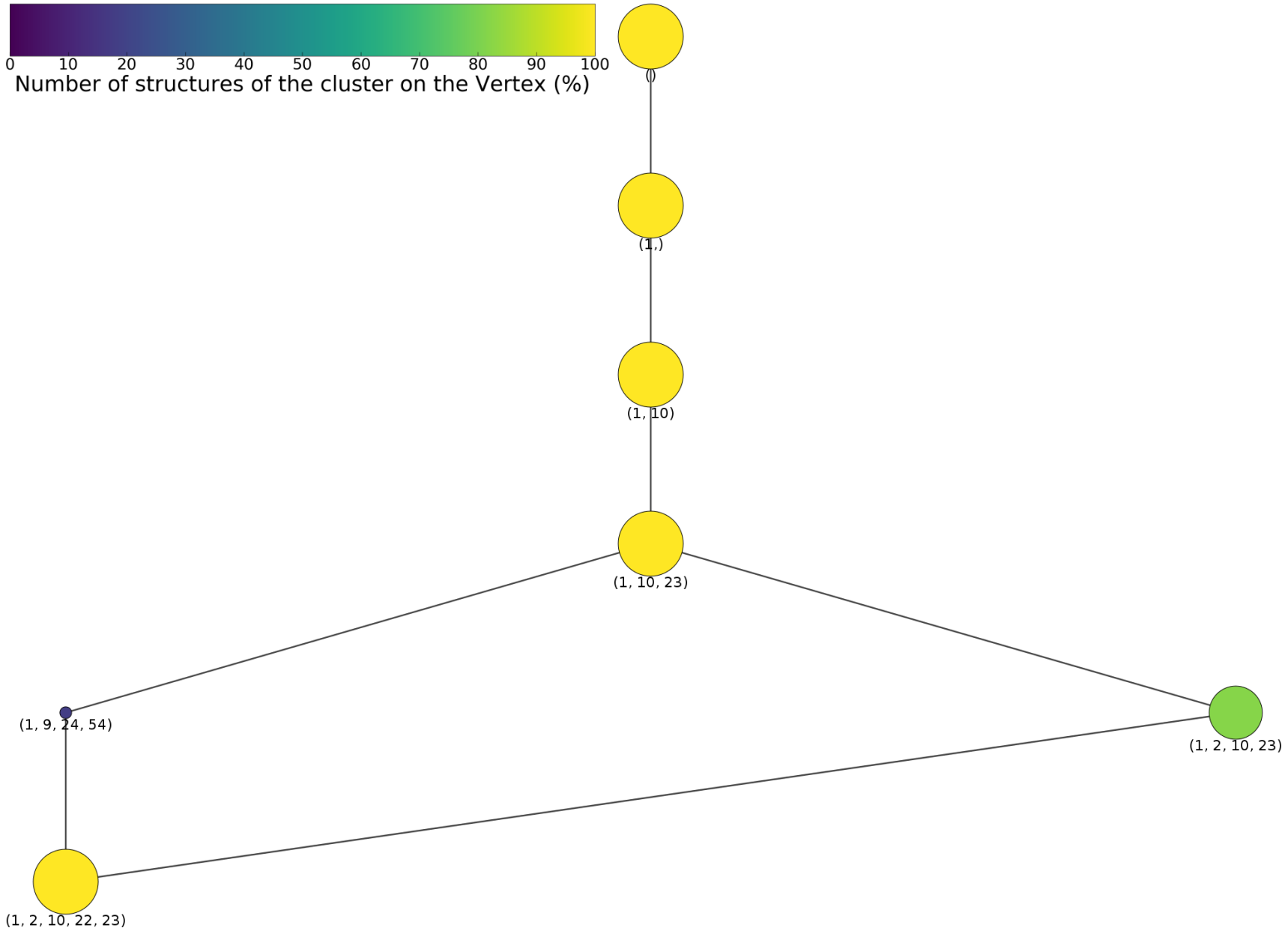

Figure D.3: Disassembly path with the energy minimum protein combination obtained for each PDB structure of the Canine and Feline Panleukopenia Virus group in each step of the removal of one protein, with heuristic II.

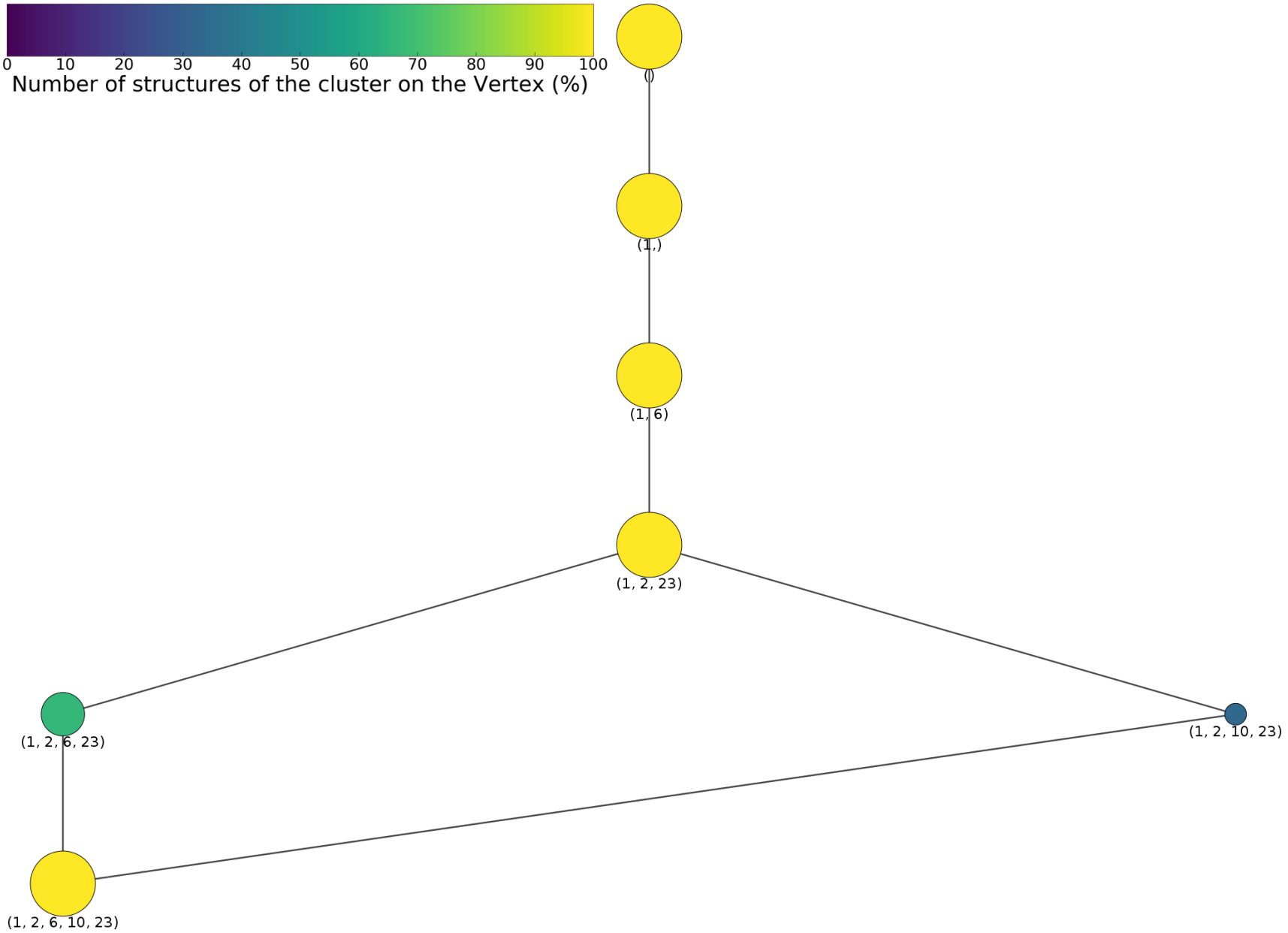

Figure D.4: Disassembly path with the energy minimum protein combination obtained for each PDB structure of the Hepatitis E Virus group in each step of the removal of one protein, with heuristic II.

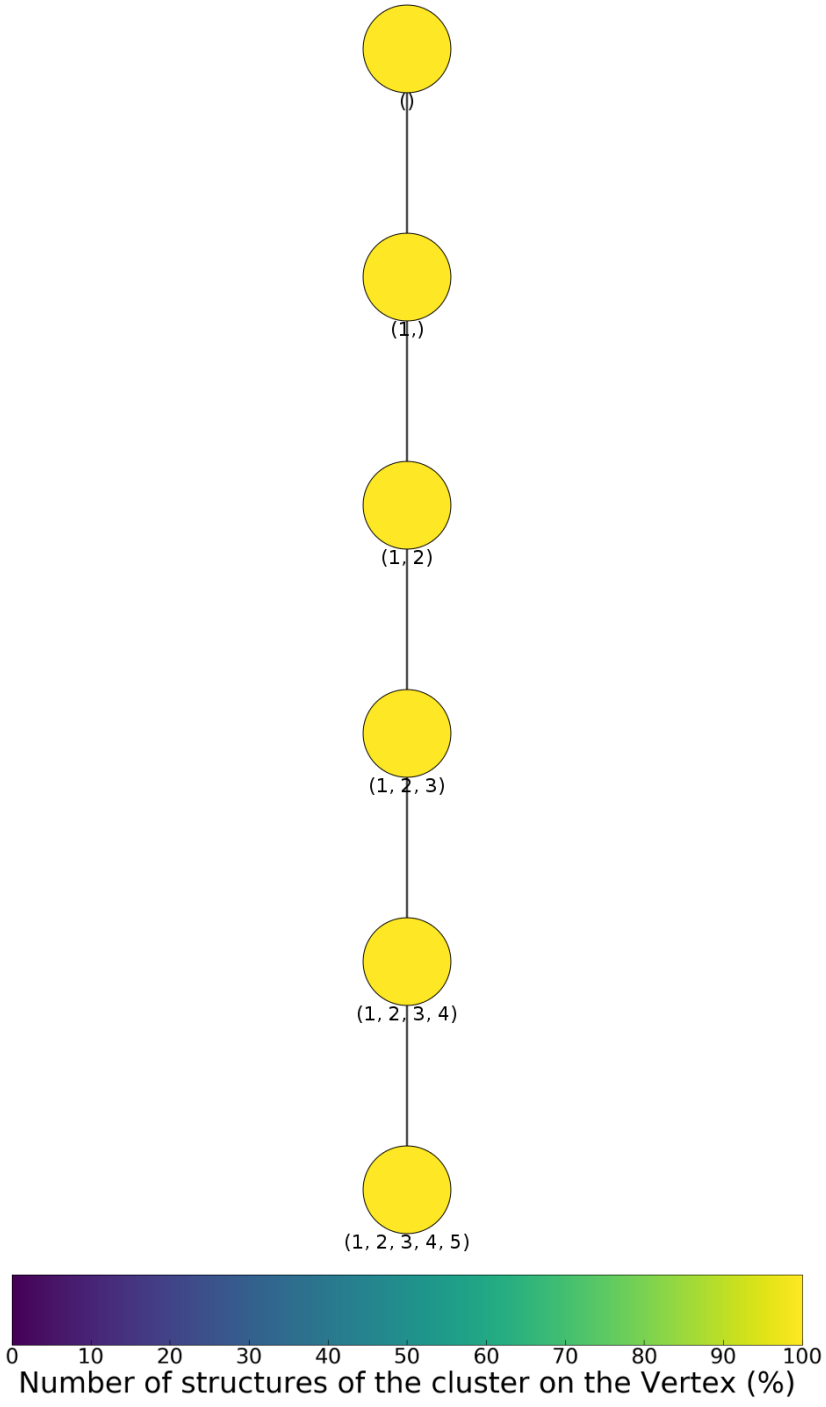

Figure D.5: Disassembly path with the energy minimum protein combination obtained for each PDB structure of the Human Adenovirus Pt-Dd group, in each step of the removal of one protein, with heuristic II.

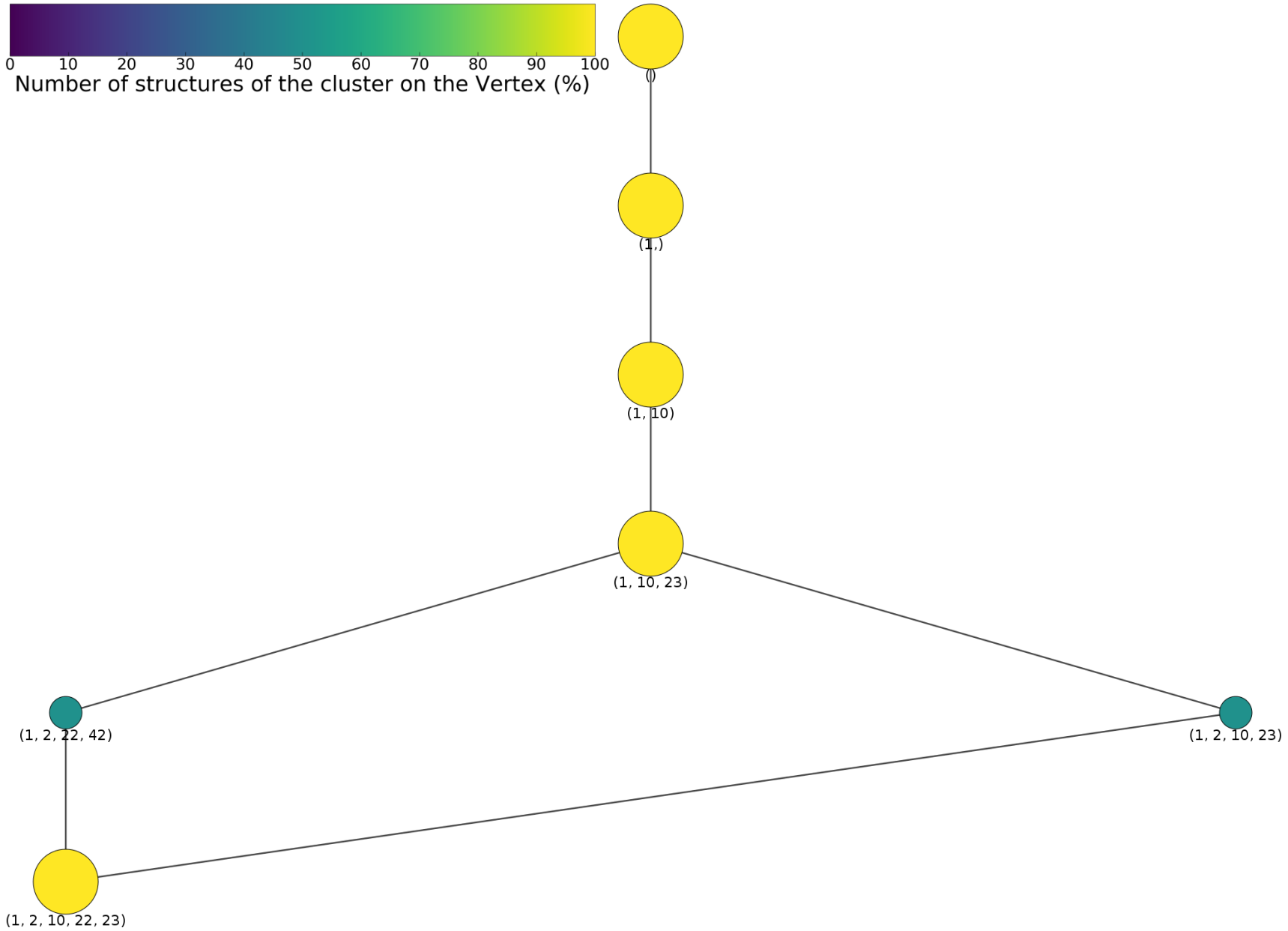

Figure D.6: Disassembly path with the energy minimum protein combination obtained for each PDB structure of the Porcine Circovirus group in each step of the removal of one protein, with heuristic II.

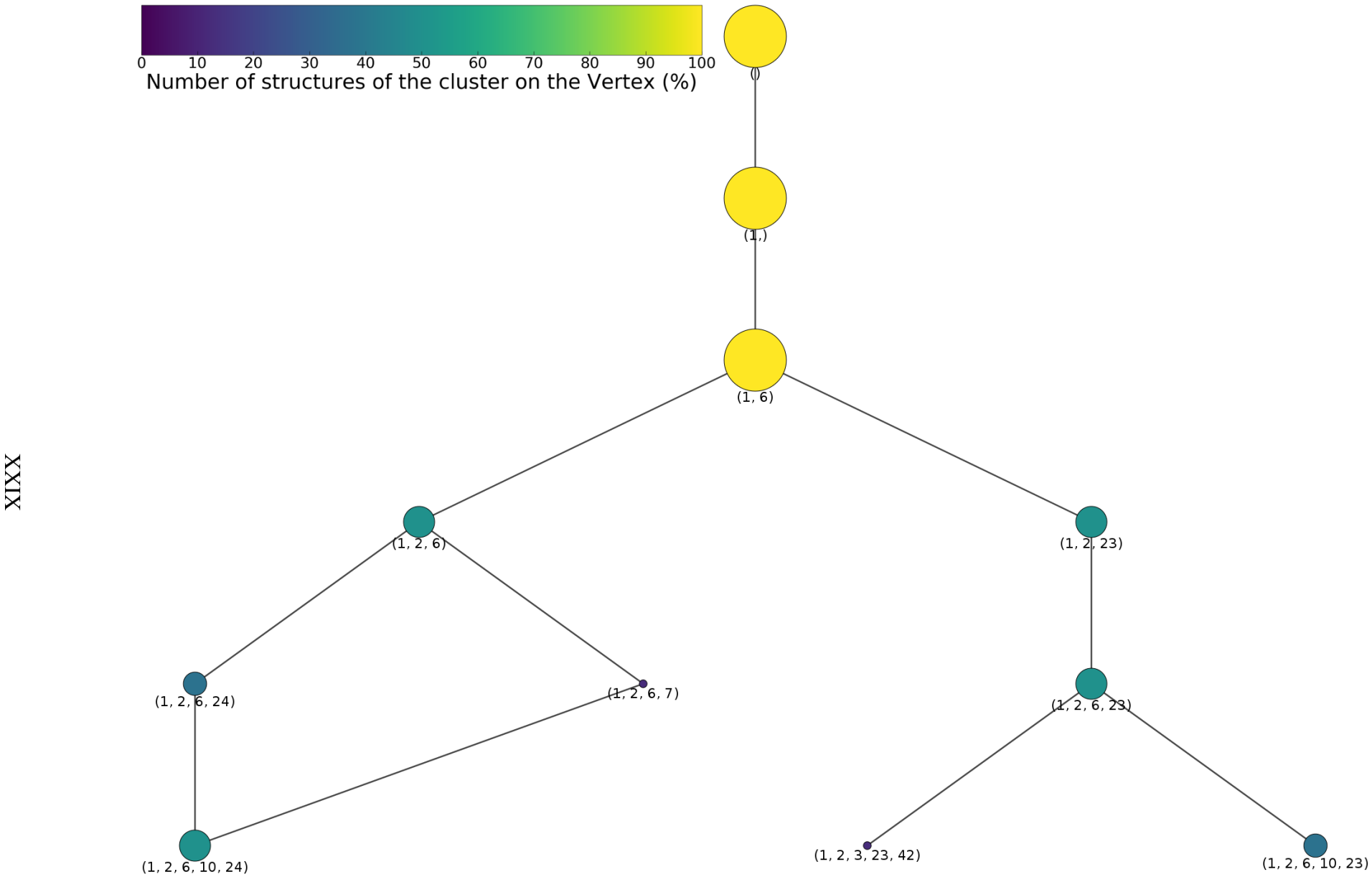

Figure D.7: Disassembly path with the energy minimum protein combination obtained for each PDB structure of the Satellite Tobacco Mosaic Virus group in each step of the removal of one protein, with heuristic II.

Appendix E

Supplementary Results using Heuristic III

Table E.1: Results when five proteins are removed, for Adeno-Associated Virus Capsid structures, using the heuristic III.

	PDB_{ID}	Energy Minimum	Second Energy Minimum
Adeno-Associated Virus	3j4p	$\{1,2,10,22,23\}$	$\{1,2,6,10,23\}$
	3ntt		$\{1,2,10,23,42\}$
	3ux1		$\{1,2,6,10,23\}$
	4iov		
	3j1q	$\{1,2,6,10,23\}$	$\{1,2,10,22,23\}$
	3ra2	$\{1,2,10,22,23\}$	$\{1,2,6,10,23\}$
	3ra8	$\{1,2,6,10,23\}$	$\{1,2,10,22,23\}$
	2g8g	$\{1,6,13,45,46\}$ $\{1,6,25,26,53\}$	$\{1,2,6,10,23\}$
	1lp3	$\{1,2,6,10,23\}$	$\{1,2,10,22,23\}$
	2qa0		
	3ra4		
	3ra9		
	3raa		
	4rso		
	5egc		$\{1,2,3,23,42\}$

Table E.2: Results when five proteins are removed, for Bovine, Human and Porcine Parvoviruses Capsid structures, using the heuristic III.

	PDB_{ID}	Energy Minimum	Second Energy Minimum
Bovine Parvovirus	4qc8	$\{1,2,10,22,23\}$	$\{1,2,10,23,42\}$
Human Parvovirus	1s58		
Porcine Parvovirus	1k3v		

Table E.3: Results when four ($N = 4$) and five ($N = 5$) proteins are removed, for *Bombyx mori*, *Galleria mellonella* and *Penaeus stylirostris* Densoviruses Capsid structures, using the heuristic III.

		Energy Minimum	Second Energy Minimum
	PDB_{ID}	$N = 4$	
Bombyx mori Densovirus	3p0s	$\{1,2,6,23\}$	$\{1,2,22,23\}$
Galleria mellonella Densovirus	1dnv	$\{1,2,22,23\}$	$\{1,2,23,42\}$
Penaeus stylirostris Densovirus	3n7x		$\{1,2,10,23\}$
		$N = 5$	
Bombyx mori Densovirus	3p0s	$\{1,2,6,10,23\}$	$\{1,2,6,10,24\}$
Galleria mellonella Densovirus	1dnv	$\{1,2,3,23,42\}$	$\{1,2,3,22,23\}$
Penaeus stylirostris Densovirus	3n7x	$\{1,2,10,22,23\}$	$\{1,2,3,41,42\}$

Table E.4: Results when five proteins are removed, for Hepatitis E Virus Capsid structures, using the heuristic III.

	PDB_{ID}	Energy Minimum	Second Energy Minimum
Hepatitis E Virus	2ztn	$\{1,2,6,10,23\}$	$\{1,6,25,26,53\}$ $\{1,6,13,45,46\}$
	2zzq		$\{1,6,14,29,45\}$
	3hag	$\{1,2,10,22,23\}$	$\{1,2,10,23,42\}$

Table E.5: Results when five proteins are removed, for Canine and Feline Panleukopenia Virus Capsid structures, using the heuristic III.

	PDB_{ID}	Energy Minimum	Second Energy Minimum
Canine & Feline Panleukopenia Virus	1c8f	$\{1,2,10,22,23\}$	$\{1,2,10,23,42\}$
	1c8d		
	1c8e		
	1c8g		
	1fpv		
	1p5y		
	1c8h		$\{1,2,10,22,42\}$
	1ijs		
	1p5w		
	2cas		
	4dpv		

Table E.6: Results when five proteins are removed, for Rodent Protoparvovirus Capsid structures, using the heuristic III.

	PDB_{ID}	Energy Minimum	Second Energy Minimum
Rodent Protoparvovirus	1mvm	$\{1,2,10,22,23\}$	$\{1,2,10,23,42\}$
	1z1c		$\{1,2,6,10,23\}$
	1z14		$\{1,2,10,23,42\}$
	2xgk		$\{1,2,6,10,23\}$
	4g0r		$\{1,2,10,23,42\}$
	4gbt		$\{1,2,6,10,23\}$

Table E.7: Results when five proteins are removed, for Satellite Tobacco Mosaic Virus Capsid structures, using the heuristic III.

	PDB_{ID}	Energy Minimum	Second Energy Minimum
Satellite Tobacco Mosaic Virus	1a34	$\{1,2,3,23,42\}$ $\{1,2,6,10,23\}$	$\{1,2,6,22,23\}$
	2buk	$\{1,2,6,10,24\}$	$\{1,2,6,11,16\}$ $\{1,2,6,31,36\}$ $\{1,2,6,21,26\}$
	4bcu	$\{1,2,3,23,42\}$	$\{1,2,7,23,33\}$
	4oq8	$\{1,2,10,22,23\}$	$\{1,2,6,10,23\}$ $\{1,2,3,41,42\}$ $\{1,2,3,22,23\}$

Table E.8: Results when five proteins are removed, for Avian Birnavirus Capsid structures, using the heuristic III.

	PDB_{ID}	Energy Minimum	Second Energy Minimum
Avian Birnavirus	1cwd	$\{1,2,10,22,23\}$	$\{1,2,10,23,42\}$

Table E.9: Results when five proteins are removed, for Porcine Circovirus Capsid structures, using the heuristic III.

	PDB_{ID}	Energy Minimum	Second Energy Minimum
Porcine Circovirus	3jci	$\{1,7,23,32,57\}$	$\{1,2,10,22,23\}$
	3r0r	$\{1,2,10,22,23\}$	$\{1,7,23,32,57\}$

Table E.10: Results when five proteins are removed, for Human Adenovirus Pt-Dd Capsid structures, using the heuristic III.

	PDB_{ID}	Energy Minimum	Second Energy Minimum
Human Adenovirus Pt-Dd	1x9t	$\{1,2,3,4,5\}$	$\{1,2,3,23,24\}$
	4aqq		$\{1,2,3,34,35\}$
	4ar2		$\{1,2,3,6,7\}$

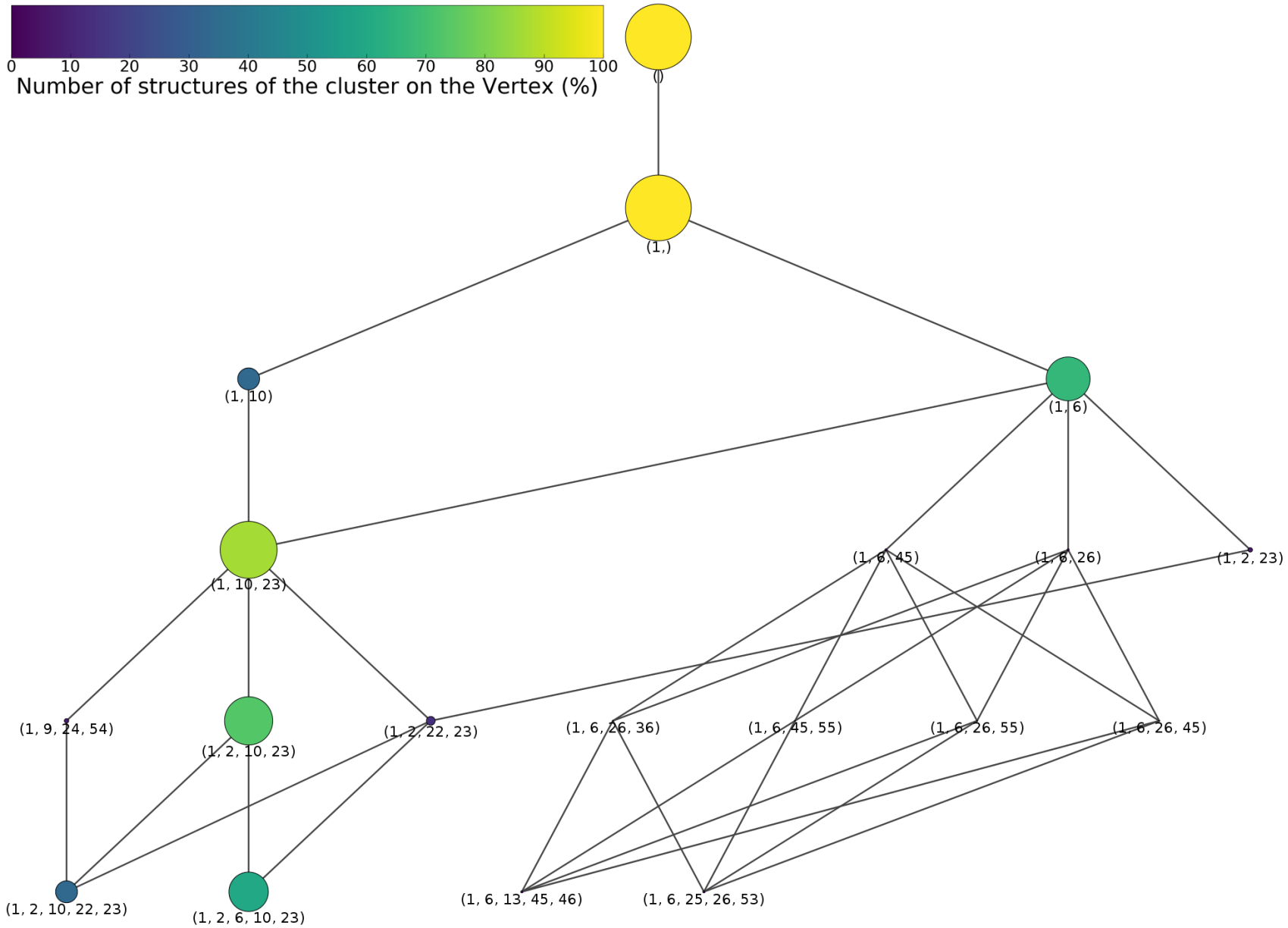

Figure E.1: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Adeno-Associated Virus group in each step of the removal of one protein, with heuristic III.

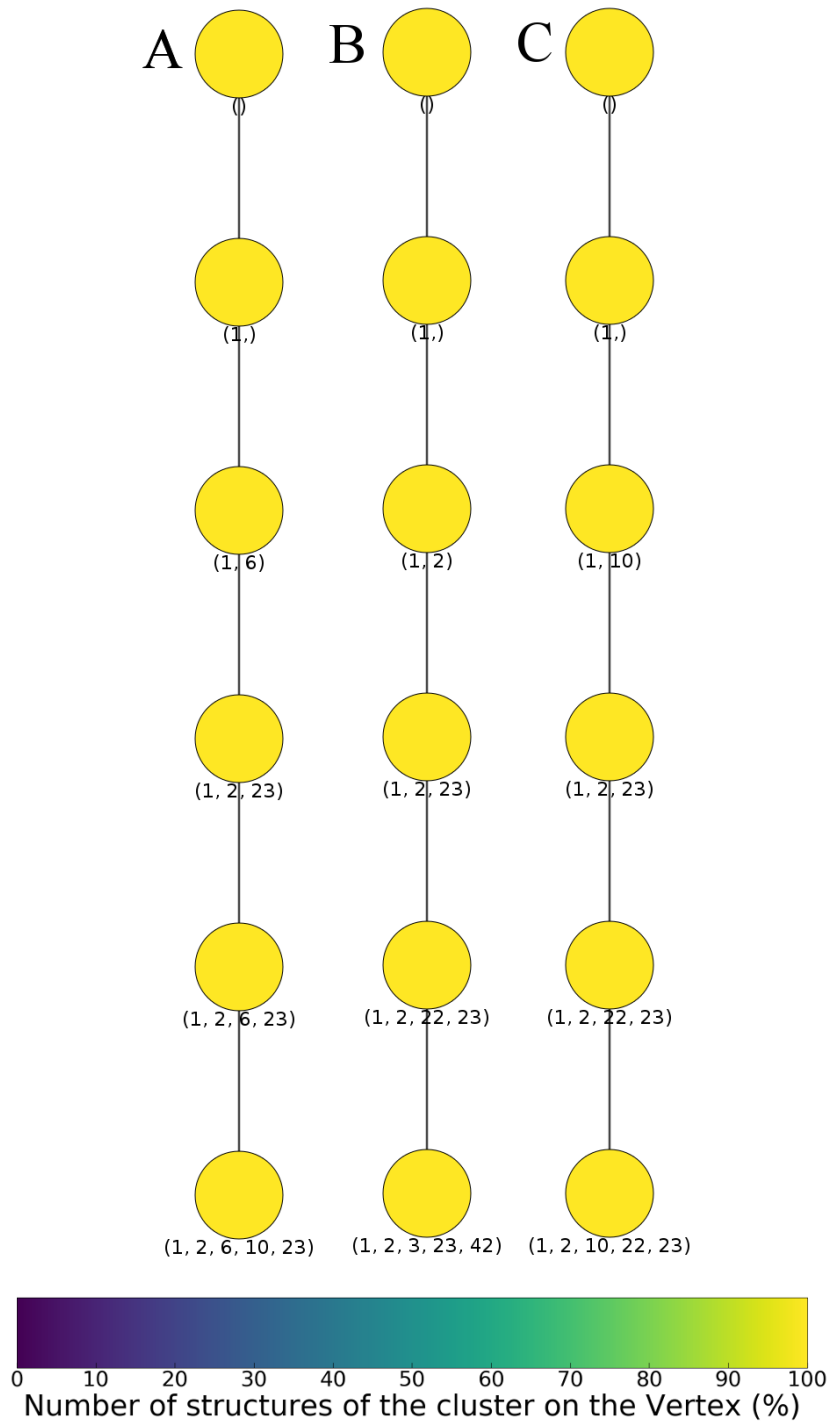

Figure E.2: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the *Bombyx mori* (A), *Galleria mellonella* (B), *Penaeus stylirostris* (C) Densoviruses groups, in each step of the removal of one protein, with heuristic III.

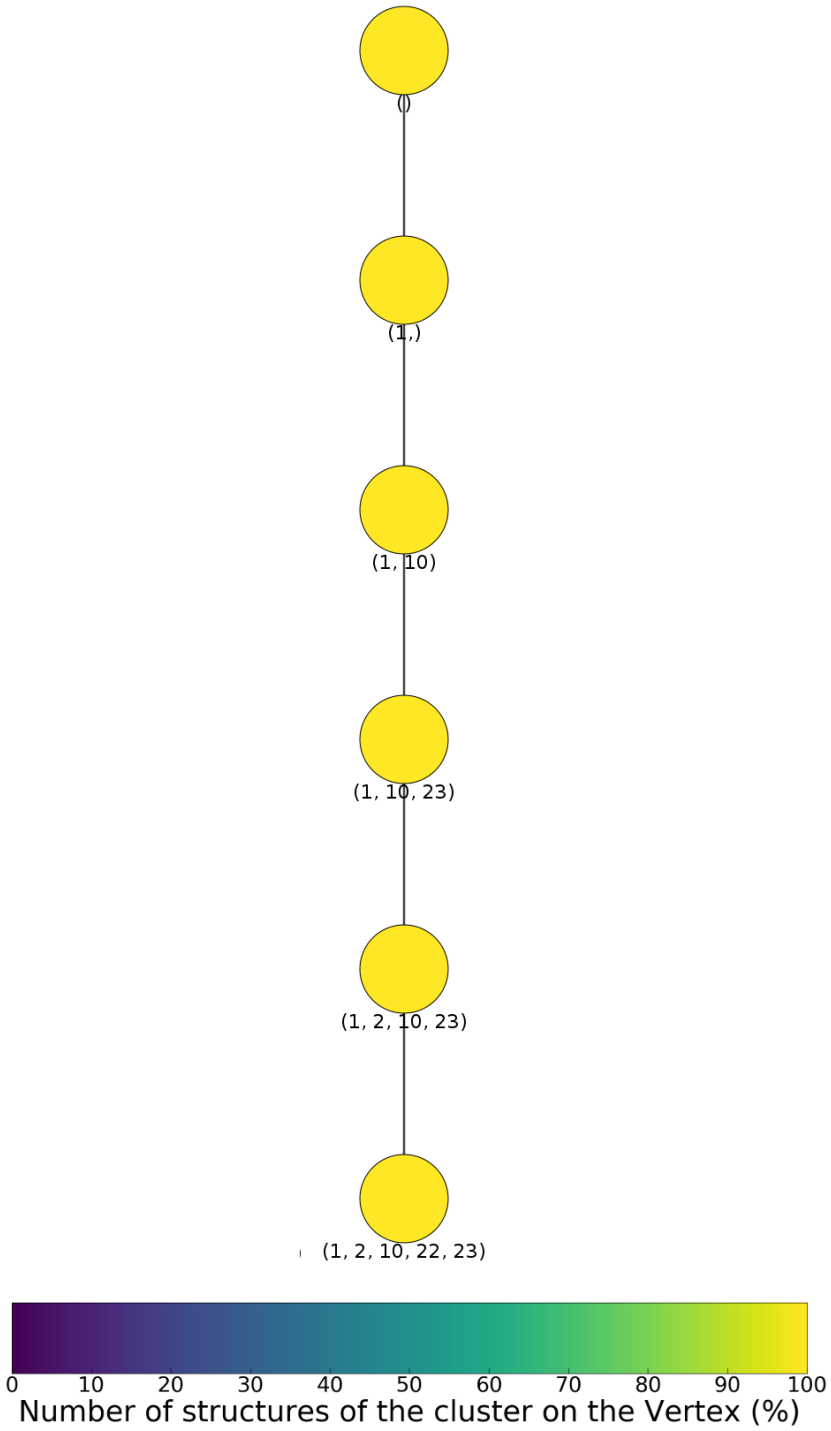

Figure E.3: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Avian Birnavirus, Bovine Parvovirus, Human Parvovirus, Porcine Parvovirus and Rodent Protoparvovirus groups, in each step of the removal of one protein, with heuristic III.

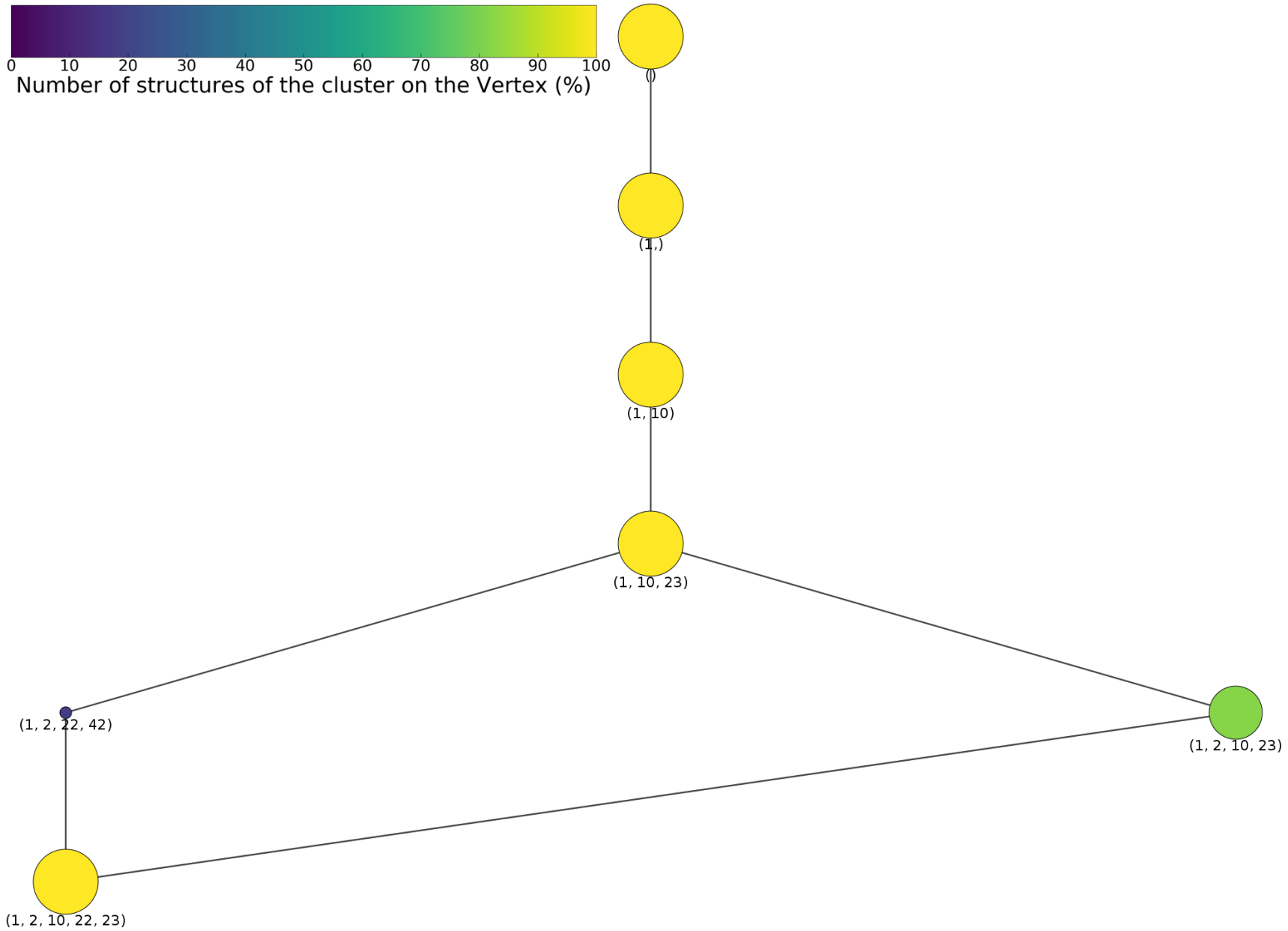

Figure E.4: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Canine and Feline Panleukopenia Virus group in each step of the removal of one protein, with heuristic III.

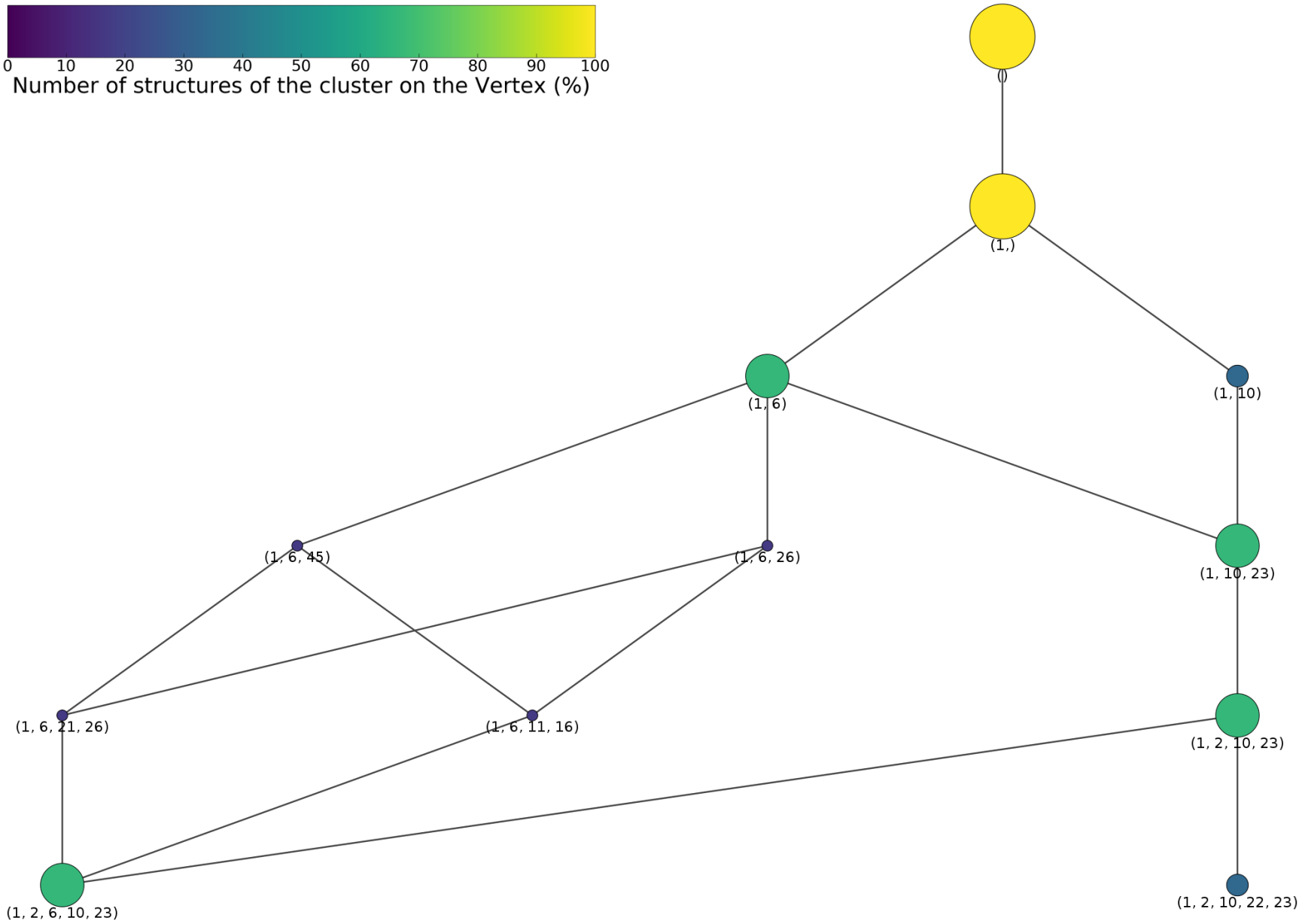

Figure E.5: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Hepatitis E Virus group in each step of the removal of one protein, with heuristic III.

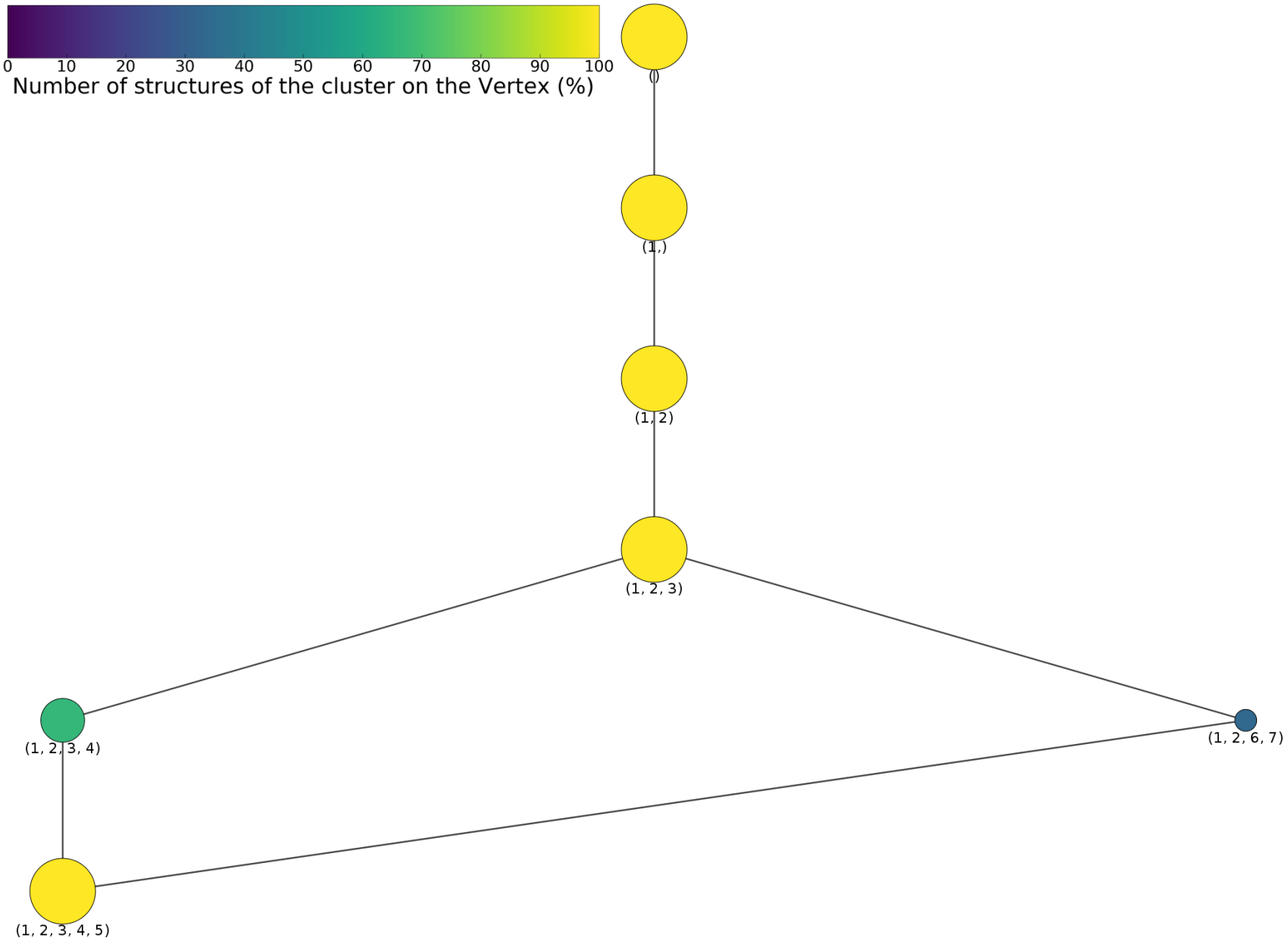

Figure E.6: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Human Adenovirus Pt-Dd group in each step of the removal of one protein, with heuristic III.

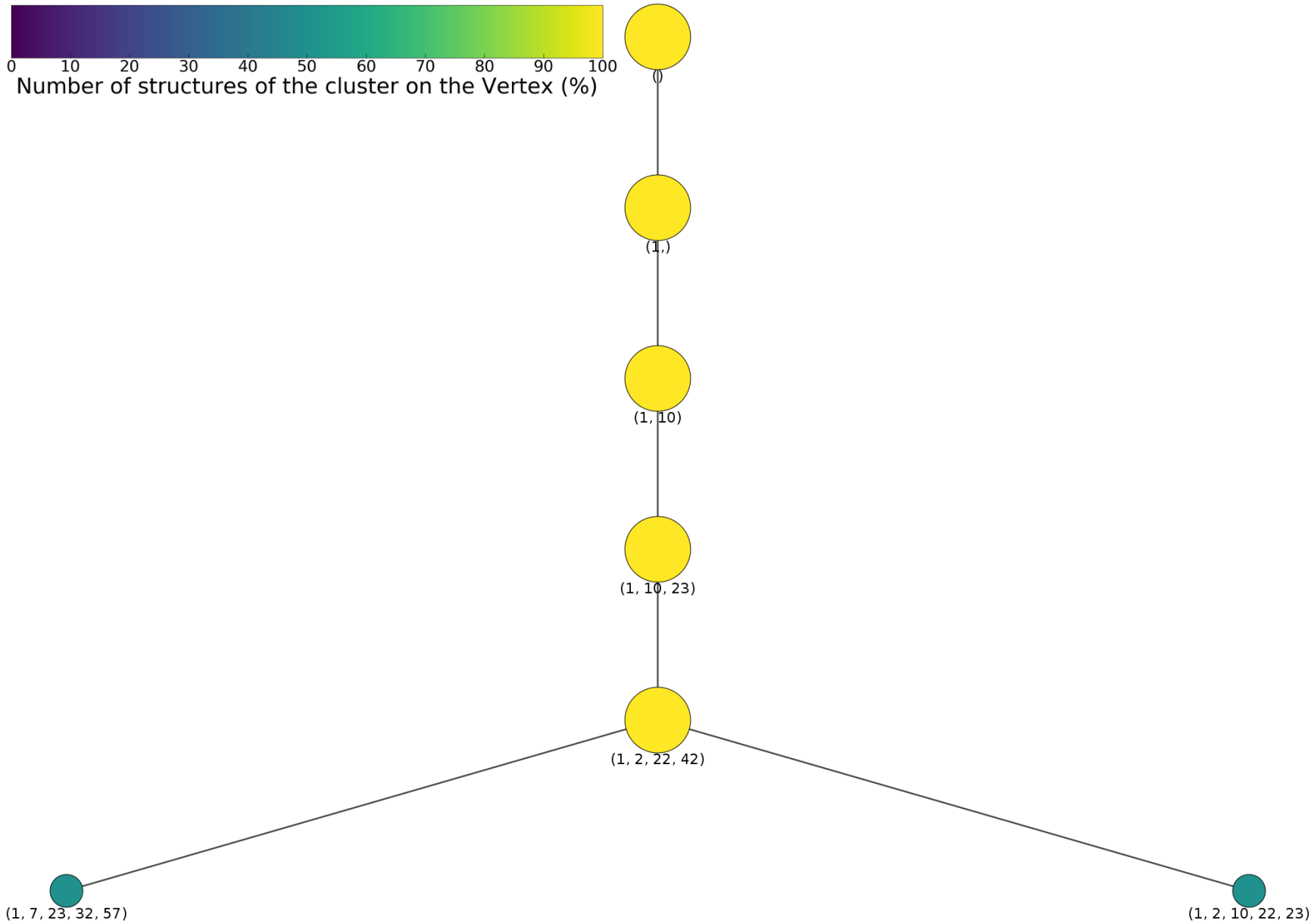

Figure E.7: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Porcine Circovirus group in each step of the removal of one protein, with heuristic III.

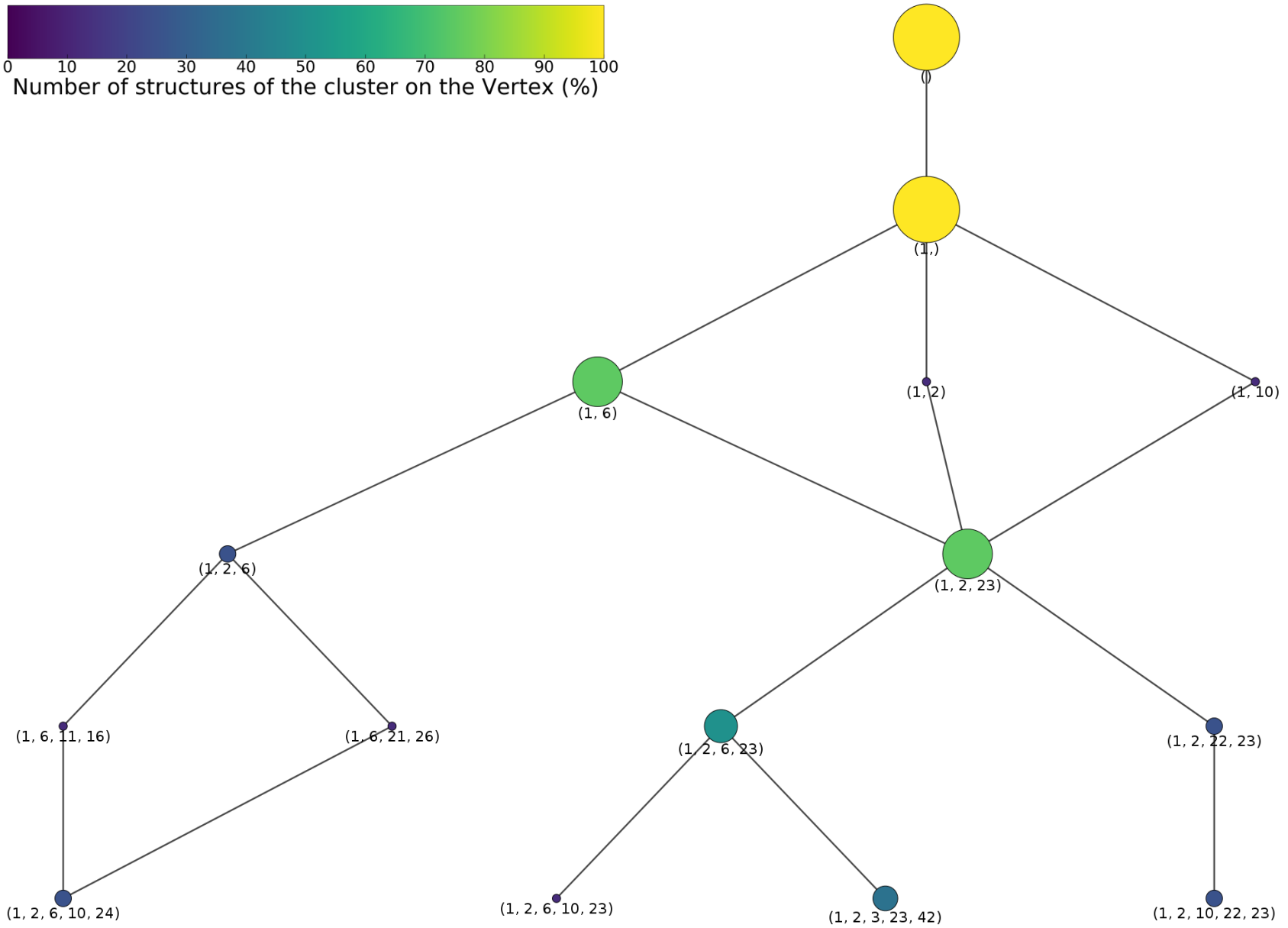

Figure E.8: Disassembly path with the energy minimum protein combination obtained for each *PDB* structure of the Satellite Tobacco Mosaic Virus group in each step of the removal of one protein, with heuristic III.