Electronic Supplementary Information

De novo Sequencing and Native Mass Spectrometry Revealed Hetero-Association of Dirigent Protein Homologs and Potential Interacting Proteins in *Forsythia* × *intermedia*

Mowei Zhou,[1] * Joseph A. Laureanti,[2] Callum J. Bell,[3] Mi Kwon,[4] Qingyan Meng,[4] Irina V. Novikova,[1] Dennis G. Thomas,[5] Carrie D. Nicora,[5] Ryan L. Sontag,[5] Diana L. Bedgar,[4] Isabelle O'Bryon,[6] Eric D. Merkley,[6] Bojana Ginovska,[2] John R. Cort,[4,5] Laurence B. Davin, and [4] Norman G. Lewis[4]

1. Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA, USA
2. Physical and Computational Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA, USA
3. National Center for Genome Resources, Santa Fe, NM, USA
4. Institute of Biological Chemistry, Washington State University, Pullman, WA, USA
5. Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA
6. Chemical and Biological Signatures Group, Pacific Northwest National Laboratory, Richland, WA, USA

Corresponding email: mowei.zhou@pnnl.gov

List of Contents for Supplemental Tables

List of Contents for Supplemental Figures

Table S1. Previously sequenced *Forsythia* × *intermedia* proteins by cDNA.

Sequence
>FiLaccase|laccase MKFSLLHLIGFLLLGGVLVLPLHAALITRHRFVLTDTPFTRLCSNKSIFVVNGQFPGPTIYAT EGDTIIVDVINQPSENVTIHWHGVKQPRYPWSDGPNYITQCPIQPGANFSQKIILSDEIGTLW WHAHSDWSRATVHGAIVIRPKNNSNYPFRTPDAEATIILGEWWKSDIRAVQNEFLGNGGDANV SDAFLINGQPGDLYPCSRSDTYNLTVESGKTYLIRMINAVMNTIMFFSIANHSVTVVGSDAAY TKPLKSDYITISPGQTIDFLLQANQTPSHYYMAARAYAVAGNFDNTTTTAIIRYKGNYTAPSS PSFPNLPGFNDTNASVNFTYRLRSLGNKNYPVDVPKNVTDKLLFTFSINLTPCPNNSCAGPFN ERFRASVNNITFVPPTIAILQAYYQRIRNVYSNNFPSNPPFTFNYTSDIIPRDLWRPQNGTEV KVLKYNSTVEIVFQGTNILAGIDHPIHLHGQSFYVVGWGLGNFNNATDPLNYNLVDPPLMNTI AVPVSGWTAVRFKASNPGVWLLHCHLERHLSWGMDMVFITQNGEGKNERILPPPPDMPPC
>AAF25357.1|FiDir1 dirigent protein MVSKTQIVALFLCFLTSTSSATYGRKPRPRRPCKELVFYFHDVLFKGNNYHNATSAIVGSPQW GNKTAMAVPFNYGDLVVFDDPITLDNNLHSPPVGRAQGMYFYDQKNTYNAWLGFSFLFNSTKY VGTLNFAGADPLLNKTRDISVIGGTGDFFMARGVATLMTDAFEGDVYFRLRVDINLYECW
>AAF25358.1|FiDir2 dirigent protein MAAKTQTTALFLCLLICISAVYGHKTRSRRPCKELVFFFHDILYLGYNRNNATAVIVASPQWG NKTAMAKPFNFGDLVVFDDPITLDNNLHSPPVGRAQGTYFYDQWSIYGAWLGFSFLFNSTDYV GTLNFAGADPLINKTRDISVIGGTGDFFMARGVATVSTDAFEGDVYFRLRVDIRLYECW

Table S2. *De novo* sequenced proteins confirmed by intact protein data (denaturing top-down or native MS) and/or bottom-up peptide data. Additional proteins were identified by bottom-up peptide data, but only selected few with high confidence (top 10 protein hits at least in one MS analysis and with >50% sequence coverage with no significant sequence gaps) are included. Some sequences determined from transcripts contained additional residues that may not be present in the protein. Annotation was based on homology. Signal peptide predictions were performed by SignalP-5.0.[1] All sequences can be found by BLAST to the recently published *Forsythia suspensa* genome[2] with near complete sequence matches (https://www.ncbi.nlm.nih.gov/assembly/GCA_013103335.1).

Sequence	Comments
>FiDir18|Dirigent protein homolog 18 kDa KTATIQIFVQDEVGGENKTVWEVARSSITADSPTLFGQVRVVDDL LTARPNKTSKKIGRVQGLITSADLKESAIAMNLNFVFTSGKYKGS TLCMLGRNPLGNAYRELAIVGGTGLFRMARGYAITSTYSYDTPTY GVLEYKIYVAYVGASTADQ	C-terminus not fully explained (-387.2 Da). See coverage map in **Figures S5** and **S7**.
>FiDir19.Fk|Dirigent protein homolog 19 kDa (based on Forsythia koreana transcript) KMTTIRVFVQDEVGGENQTVWEVARSKITADSPTLFGQVRVVDDL LTAKPNKTSKKVGRVQGLITSADLQVSAIAMSMNFIFTIGKYNGS TLCMQGRNQLGNDYRELAIVGGTGLFRMARGYAITSTGTNYIC	C-terminus not fully explained (1809.9 Da). See coverage map in **Figures S6** and **S8**. The C-terminal residues cannot be confirmed.
>FiDir19 - Dirigent protein homolog 19 kDa (based on Forsythia suspensa genome) KMTTIRVFVQDEVGGENQTVWEVARSKITADSPTLFGQVRVVDDL LTAKPNKTSKKVGRVQGLITSADLQVSAIAMSMNFIFTIGKYNGS TLCMQGRNQLGNDYRELAIVGGTGLFRMARGYAITSTYSYDTPTY GGVMNELMIHHWVVWP	See coverage map in **Figures S6** and **S8**. This sequence better explains the experimental data. The yellow highlighted region is different from FiDir19.Fk sequence shown above.
>nsLTP – nonspecific lipid transfer protein AGECGRTPINAAATSLSPCLGAATNVRVKVPPPCCAKVNALIKST PKCLCAVLLSPLAKKAGIKPGIAITIPKRCNIRNRPVGKKCGGYT VP	Top down coverage map is in **Figure S13**. Also confirmed with full peptide coverage (data not shown).
>Beta-fructofuranosidase (invertase) LFVLYCQKPMELANYKVYWVFPVFLFCFLILNNGVVEASHKVYLN LQSVSPVNVNQVHRTGYHFQPTKNWINDPNGPMYYNGIYHLFYQY NPKGAVWGNIVWAHSVSKDLINWKKVEHAIIPSKPFDQYGCWSGS ATILPGNKPVIMYTGIIDKNNTQVQNYAVPANLSDPYLRVWNKPD NNPLVVADESINKTAFRDPTTAWLGRDSHWRISLGSRRKHRGIAY LYRSRDFKNWVKAKHPLHSVAGTGNWECPDFFPVSVQGTNGLDTS VLRGNVKHVLKVSLDATRYEYYTIGTYDAKKDRYIPGKDMVDGWK GLRYDYGNFYASKSFFDPAKNRRVLWGWANESDAVMDDIAKGWAG IQLIPRTIVLDPSGKQLLQWPIEELETLRGNRVELRNNTLEKGER LEIKGITVAQADVEVIFSFGSLDKAEPFDPSWDRYDAQKLCSQKG STVQGGVGPFGLLTLASENLEEYTPVFFRIFKDQDKHIVLMCSDA TRSSLADKSGTYRPSFAGFVDVDLADKKLSLRSLIDHSVVESFGA RGKTCITSRVYPTLAIYENAHLYAFNNGTETVKIECLKAWSMERP GLMNH	Likely the major species in the ~60 kDa gel band. Full sequence listed is based on the transcript. Sequence with mass spectrometry coverage is shown in red. The full peptide coverage map is in **Figure S14**.
>Peroxidase 4 FFYSKLINPSMATFSVVFFIFVMLFIGSSSAQLSTNFYDKTCPKV LTTVNSVVRSAVAKEKRMGASLLRLHFHDCFVQGCDASVLLDDTS SFTGEKTAGPNNNSLRGFNVVDNIKSKVEAVCPGVVSCADILAIA ARDSVVILGGPSWKVKLGRRDSKSASFSAANSGVIPPPTSTLNNL RNRFKARGLSTKDMVVLSGAHTIGQARCTSFRVRIYNESNIDTSF	Full sequence listed is based on the transcript. Sequence with mass spectrometry coverage is shown in red. The starting residue also matches to signal

ARTRQRKCPLTVGSGDNNLAPLDVQTPTAFDNDYYKNLINKKGLL HSDQILYNGGSTDSLIESYSKNSNSFNSDFAAAMIRMGNISPLTG SSGQIRKNCRRPN	peptide prediction. The full peptide coverage map is in **Figure S15**.
>Basic blue protein-like ATYVVGGKGGWTFNVDSWPNGKRFKAGDTLVFNYNSALHNVVAVN KGGYQGCTTPRRAKVYKTGKDQIKLVKGQNFFICNFAGHCQSGMK ISINA	Top down coverage map is in **Figure S16**.
>Germin-like protein 1 RVSADPDLLQDLCVADLTSAVKVNGFACKSNVTEEDFFFTGLAKP GATNNSMGSVVTAANVQKIPGLNTLGVSLARIDYAPGGLNPPHTH PRATEIVFVLEGELDVGFITTANVLVSKSIKKGEIFVFPRGLVHF QKNNGKVAAAVIAAFNSQLPGTQSIAATLFAASPTVPDNVLTKAF QVGTKEVEKIKSKFAPKK	Full sequence listed is based on the transcript. Sequence with mass spectrometry coverage is shown in red. The full peptide coverage map is in **Figure S17**.
>Germin-like protein 2 PLQDFCEATSTDHNKNNTAVVLKMAVRILISVLIISLFSFTYASD PAPLQDFCVAVKDNEAKVFVNGKICKDPNMVSADDFFFPGLNKPG NTSNAQGSKVTPVNVNQLPGLNTLGISLVRIDYAPYGLNPPHTHP RATEVLVVVEGTLFVGFVTSNPADPNVKNKLFTKTLYPGDVFVFP QGLIHFQYNVGKTDAFVFAGLSSQNPGVITIANAVFGSDPPINLD VLTKAFQVDANVIKYLQAPFM	Sequence hit in the assembled transcript data. Only a segment of sequence near the C-terminus (in red) was confirmed by native top-down in **Figure S18**.
>Gamma-glutamyl transpeptidase 1-like, glutathione hydrolase 1-like EGSSEVGFLHKQFMASVLMLNSASILLLFFSILCLSSSLSTGLAS TRREVIRANNGVVATDHGQCSTIGRNVLLEGGHAVDAAVAAALCL GVVSPASSGIGGGAFMLVRSADGLTKAFDMRETAPNKASENMYAG NVVLKSGGALSVAVPGELAGLHEAWEQYGKISWDRLVRPAAQLAH NGFKISPYLHMQMVKTESGIMTDKGLRGIFTSNGSLLQPGDIIYN RKLAKTLKAISKYGVKALYNGTIGFNLVKDVRKAGGILTMRDLQH YRVKLREPISVDVMGVRILAMPPPSGGGAAMSLILNILAQYEGLL NISDSLVIHREIEALKHAISMRMNLGDPDFVNITDVFKDMLSTKF AAKLRKTIFDNMTFNASHYGGKWNQIHDHGTSHISIVDGKRNAVS MTTTINSYFGSKFLSPTTGILLNNEMDDFSMPAKNMENIPPPAPA NFIHPGKRPLSSMNPAIVLKGKKLKAVIGASGGSLIIAGTTEVFL NYFARGMDPLSSVMAPRSYAQLIPNVLQYENWTAVTGDHFEVPQT TRDALKKRGHVLQSLFGGTICQFVVQEELDSSTSRKLVAVSDPRK GGFPAGF	Regions with peptide coverage shown in red. Gray shaded region was detected in top down. The full-length form was likely in the ~60 kDa gel band.
>Pollen Ole e 1 allergen and extensin family KPPSHPPVKPPSHPPAKPPSPPPAKPPTLPPSHPPIRKAVGVQGV VYCKSCKYRGIGTLLGASPLAGAVVKLQCNNTKWGLVEQTKTDKN GYFFFKPSKLTTAAFHKCKVFLISSPLVTCSVPTNLGSGASGAIL IPSLKPPVKPLPFQLFTVGPFAFEPAKKVPCHH	Regions with peptide coverage shown in red. Gray shaded region was detected in top down.
>Subtilase family IMPPCLKPRRHSCKFLSSPTLFPHFSHTHKLSIFSFSKTMGLGSG TVIWLFSIALVLQSCIFTVSAKKIYIVRMKHHQMPTSYSTQSDWY ADHLQSLTSATPDSLLYTYEAAYHGFAVALDAEEAESLRQSDSVL GVYEDTIYNLHTTRTPEFLGIERELGLWAGHGPQELNQASQDVIV GVLDTGVWPESKSFEYADMPEVPSRWRGQCEVADDFDPKINCNKK LIGARFFSRGHNMASGGKELQSPRDQDGHGTHTASTAVGYYVAKA NLLGYASGTARGMATHARVATYKVCWETGCFGSDILAGMERAILD GVDILSLSLGGGSVPYYRDTIAIGAFAAMEKGILVSCSAGNGGPA KATLANVAPWIMTVGAGTLDRDFPAFATLGDGQKFSGVSLYSGKG	Likely in the ~60 kDa gel band. Regions with peptide coverage shown in red.

MGEKLVGLVYNKGGNISSNLCLAGSLDPTIVRGKVVLCDRGISAR AEKGVVVRDAGGVGMILANTAAMGEELVADSHLLPAVAVGRKVGD IIRKYAGTDKNPTAVLRFGGTVVNVKPSPVVAAFSSRGPNMVTPQ ILKPDVIGPGVNILAAWSEAVGPTGLQQDTRRTKFNIMSGTSMSC PHISGLAALLKAAHPDWSPSAIRSALMTTAYTLDNTNSPLRDAVD NSFSNPWAHGSGHVDPRKALSPGLVYDITPEEYIMFLCSLDYTIE HVQTIVKRSNVTCSKKFTDPSQLNYPSFSVLFGKSRVARYSRELT NVGAARSIYQVAVEAPKNVVVTVKPSSLVFRNVGDKQRYTVTFVS KKGVNQLGRNAFGSISWKNAQHLVKSPVAFSWTHI	
>Purple acid phosphatase DILGSGAQDKEMGTPGTRFSWFSVLGFVLNAAILCNGGITSSFVR KVEKTMDMPLDSDVFRVPPGYNAPQQVHITQGDHLGKAVIVSWVT VDEPGSNTVLYWSESSKDKKEAKGKLTKYKYFNYTSGYIHHCTIK NLEYTTKYYYEVGIGHTTRTFWFTTPPEVGPDVPYTFGLIGDLGQ SYDSNKTLTHYEKNPTKGQTLLFVGDLSYADNYPNHDNVRWDTWG RFVERSLAYQPWIWTVGNHEIDFAPEIGETKPFKPYSHRYHTPYK ASDSTSPFWYSIKRASAYIIVLSSYSAYGKYTPQYKWLEQELPKV NRSETPWLIVLMHSPWYNSYNYHFLEGETMRVMYEPWFVKYKVDV VFAGHVHAYERSERVSNIAYNIVNGLCTPVPDQSAPVYITIGDGG NLEGLATNMTEPQPKYSAYREASFGHATLDIKNRTHAYYSWHRNQ DGYAVEADTMWFFNRFWHPVDDSTTAES	Likely in the ~60 kDa gel band. Regions with peptide coverage shown in red.
>Eukaryotic aspartyl protease family LNSSFLQSCMGGKLLISLAILFFSSVALGITPNCNIPEQGSTIQV IHVNSPCSPFRSKTHLSWEDTVLQMQSADKERLIYLSSLVAGRSI VPVASGRQITQNPTYILRAKFGTPPQTLLMAMDTSSDAAWIPCSG CAGCGATAFDTAKSTSFKNLSCGAAQCKQVPNPSCAGTTCGFNLT YGSSSIAASLVQDTIALATDPVPGYTFGCIQKATGSSIPAQGLLG LGRGPLSLLSQTQNLYKATFSYCLPSYKSPNFSGSLRLGPNSQPI RIKYTQLLKNPRRSSLYYVNLVGIKVGKGLVKIPPTAFAHDPNTG SGTVIDSGTVFTRLVQPAYIAVRDAFRRRMGNAVVSSLGGFDTCY TVPVTVPTISFMFSGMNMTLAQDNFLIRSAVGTTSCLAMAAAPDN VNSVLNVIASFQQQNHRILIDVPNSKLGVARETCT	Likely in the ~60 kDa gel band. Regions with peptide coverage shown in red.
>Glycosyl hydrolase superfamily EMGFFAFILLSLVALSPSFTSGEMEPKIGICYGQLGNNLPTQSKS VQLIKKLGAKRVKIYDANTKILKALKGTDLQASVMVPNEIISNIS TNQTLADEWVKTNVVPFYPKTLIRYLLVGNEILSNPPNTTWFNLV PAMRKIRWSVKKFGLKKIKVGTPLAMDALESSFPPSNGTFRSDVS EKVILPLLSFLNRTKSFFFVDVYTYFAWMNQPAQINLQYALLEPT NITYTDPVSGLTYTNLLDQMLDSVIFAMKKLGYPNIRLFIAETGW PNGGDVDQIGANIYNAATYNRNVVKKFMAKPPIGTPARPGVVIPT LIFALYNENQKPGPGTERHFGLLYPNGSYVYPIDLSGKTLKSDYP PLPKPTNNEPYKGKIWCVVAKGANRTALAGALSYACGQGNGTCDP IQPGGRCYKPNTLVKHASYAFSSYWSQFRKNGGTCYFNGLAVQTA HNPSHGSCKFPSVTL	Likely in the ~60 kDa gel band. Regions with peptide coverage shown in red.

Table S3. Information about characterized DP homologs.

Name	Known Substrate/Product	Reference
AtDir5: *Arabidopsis thaliana* dirigent protein 5	Coniferyl alcohol to (–)-pinoresinol	Kim et al. J Biol Chem. 2012, 287, 33957 DOI: 10.1074/jbc.M112.387423
AtDir6: *Arabidopsis thaliana* dirigent protein 6 PDB:5LAL	Coniferyl alcohol to (–)-pinoresinol	Gasper et al., Plant Physiology. 2016, 172, 2165. DOI: 10.1104/pp.16.01281
ESB1/AtDIR10: *Arabidopsis thaliana* Enhanced Suberin 1 protein	Unknown; inactivation disrupts Casparian strip lignin in root endodermal cells	Hosmani et al., Proc. Nat. Acad. Sci. 2013, 110, 14498. DOI: 10.1073/pnas.1308412110
AtDir12: *Arabidopsis thaliana* dirigent protein 12	8-O-4′-linked feruloylcholine- and sinapoylcholine (SC)-conjugated lignans	Yonekura-Sakakibara, et al., Plant Cell 2021, 33, 129. DOI: 10.1093/plcell/koaa014
FiDir1/2: *Forsythia* x *intermedia* dirigent protein	(+)-Pinoresinol-forming dirigent protein	Davin et al., Science 1997, 275, 362. DOI:10.1126/science.275.5298.362
GhDir4 & GhDir7: *Gossypium hirsutum* dirigent protein	Hemigossypol to (+)-gossypol	Effenberger et al., Angew Chem Int Ed. 2015, 54, 14660 DOI: 10.1002/anie.201507543
PsDRR206: *Pisum sativum* dirigent protein 206 PDB: 4REV	Coniferyl alcohol to (+)-pinoresinol	Kim et al. J Biol Chem. 2015, 290, 1308. DOI: 10.1074/jbc.M114.611780
GePTS1: *Glycyrrhiza echinate* pterocarpan synthase 1 PDB: 6OOC	(3*R*,4*R*)-DMI to (–)-medicarpin; (3*S*,4*R*)-DMI to (+)-medicarpin (DMI: 7,2′-dihydroxy-4′-methoxyisoflavanol)	Uchida et al., Plant and Cell Physiology. 2017, 58, 398 DOI: 10.1093/pcp/pcw213
GmPdh1: *Glycine max* Pdh1 **PvPdh1**: *Phaseolus vulgaris* Pdh1	Unknown but sequence similar to PsDRR206; probably pinoresinol-forming dirigent protein. Inactivating mutations cause indehiscence (pod shattering resistance) enabling domestication of soybeans	Funatsuki et al., Proc. Nat. Acad. Sci. 2015, 111, 17797. DOI: 10.1073/pnas.1417282111 Parker et al., bioRxiv 2019. DOI: 10.1101/517516
Hfr DrD: *Triticum aestivum* Hessian fly disease resistance dirigent-like protein	Unknown; A29→V and S48→L mutations confer resistance to Hessian fly larval attack in wheat	Subramanyam et al., Arthropod-Plant Interactions. 2013, 7, 389. DOI: 10.1007/s11829-013-9253-4 Tan et al., Molecular Breeding. 2015, 35:216. DOI: 10.1007/s11032-015-0410-6
PsPTS1: *Pisum sativum* pterocarpan synthase 1 PDB: 6OOD	(3*R*,4*R*)-DMI to (–)-medicarpin; (3*S*,4*R*)-DMI to (+)-medicarpin	Meng et al., J Biol Chem. 2020, 295, 11584 DOI: 10.1074/jbc.RA120.012444
PsPTS2: *Pisum sativum* pterocarpan synthase 2	Under investigation	

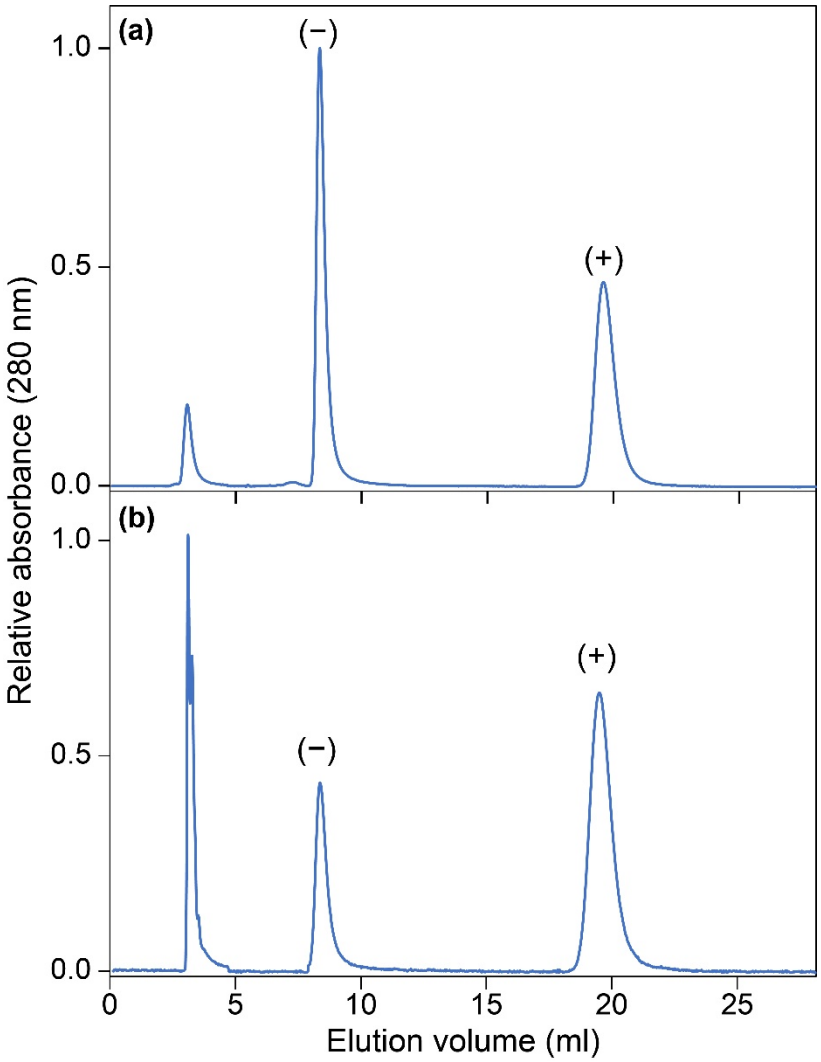


Figure S1. Chiral HPLC analyses of pinoresinol antipodes showing separation of the stereo-isomers of (a) racemic standard, and (b) formed after incubation of coniferyl alcohol with MonoS fractions eluted with 333 mM Na_2SO_4 of the *Forsythia* extract. The first peak at ~3min is void volume. The ~8 min peak is (–)-pinoresinol, and the ~20 min peak is (+)-pinoresinol. The MonoS fractions containing the active dirigent proteins promoted the formation of (+)-pinoresinol.

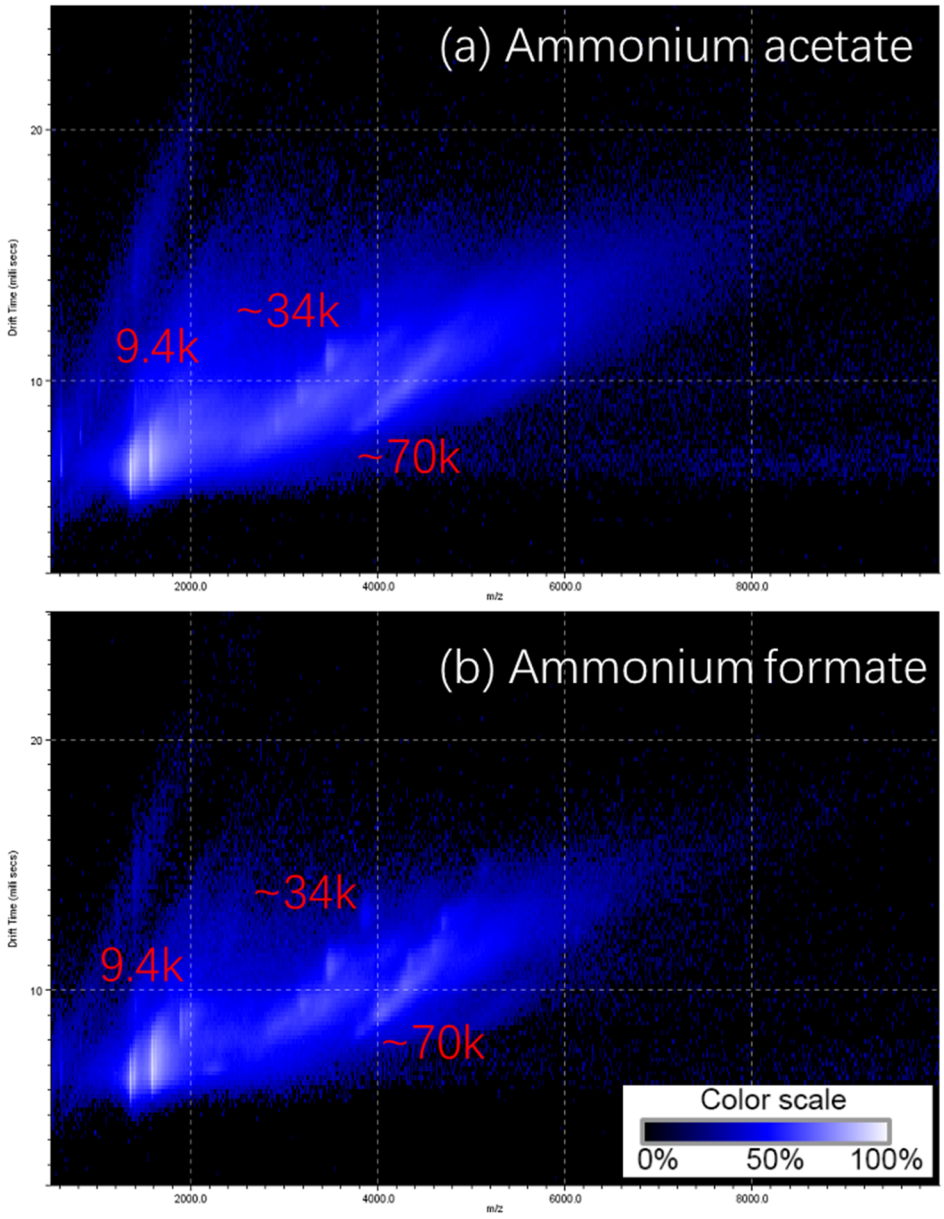


Figure S2. Representative native MS spectra of MonoS fraction in (a) ammonium acetate, and (b) ammonium formate. The x axis shows the *m/z*, the y axis shows the drift time in milliseconds by ion mobility. The color represents the relative intensity, with color scale (after log transform) shown at the bottom of the second spectrum. The major species detected were in the same *m/z* range at different pH values.

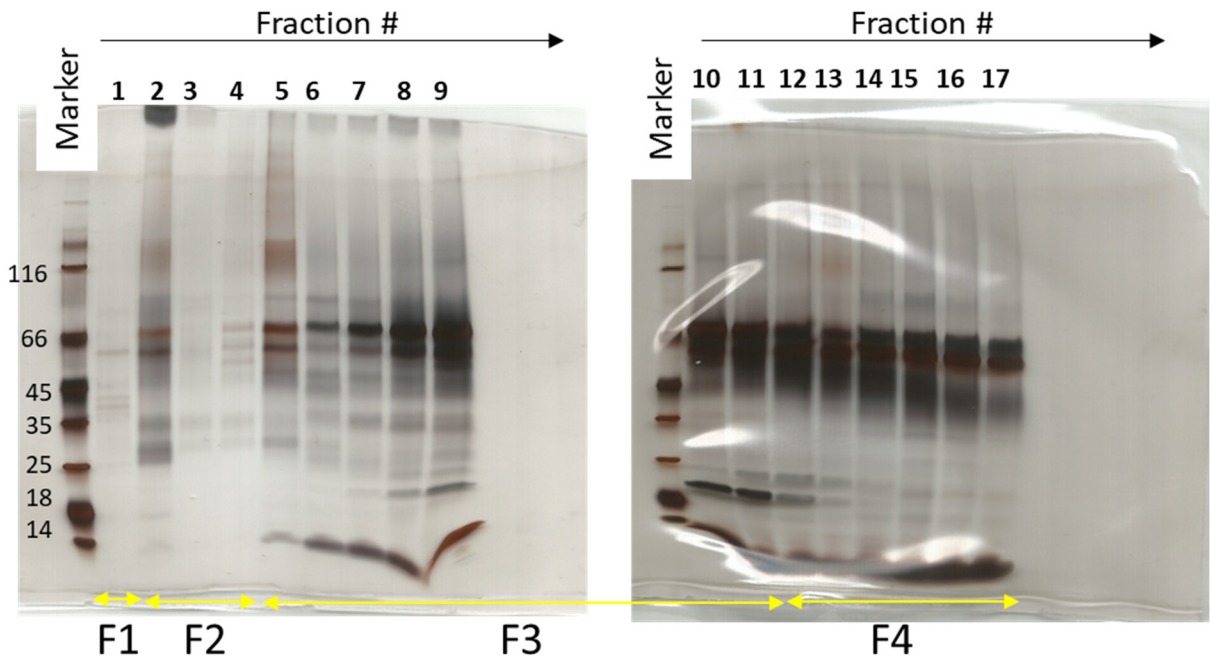


Figure S3. Representative denaturing gel images of second cation exchange fractionation with PoroS column containing *Forsythia* × *intermedia* DPs. Fractions were pooled as shown by the arrows into 4 (F1-F4) for MS analysis.

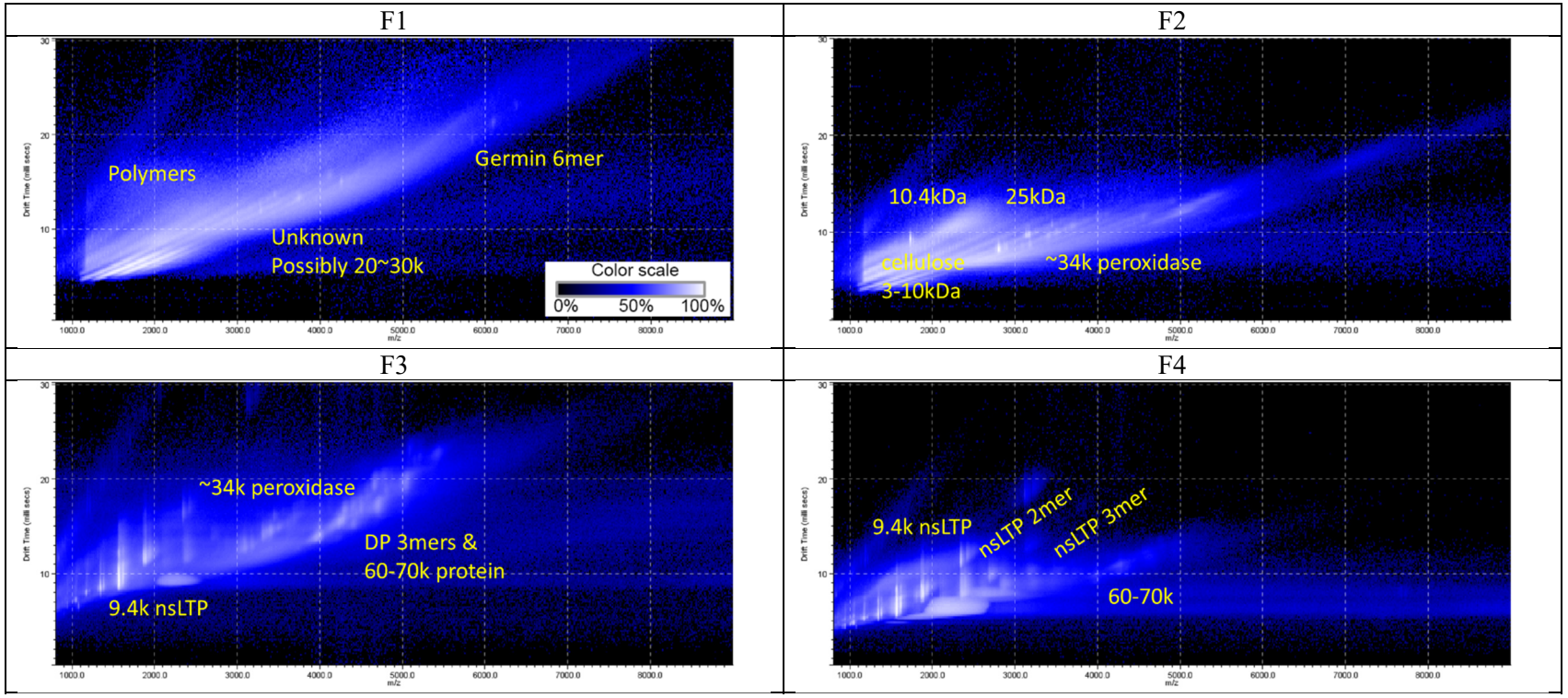


Figure S4. Native MS spectra for pooled fractions F1-F4 shown in Figure S3. The x axis shows the *m/z*, the y axis shows the drift time in milliseconds by ion mobility. The color represents the relative intensity, with color scale (after log transform) shown at the bottom of the first spectrum. Assignments are overlaid with the peaks in yellow text. Species that were confidently identified are named, other unknowns are labeled with molecular weight in Da. Germin and a 10.4 kDa protein were also detected at relatively low abundance (not easily detectable in the initial fraction as shown in Figure 1). We also detected some low abundance unknown complexes > *m/z* 4000 in F2, but it was difficult to identify their protein subunits (data not shown). While the pinoresinol forming DPs (FiDir1/FiDir2) were below the detection limit using native MS, bottom-up data showed several peptide hits (Figure S5).

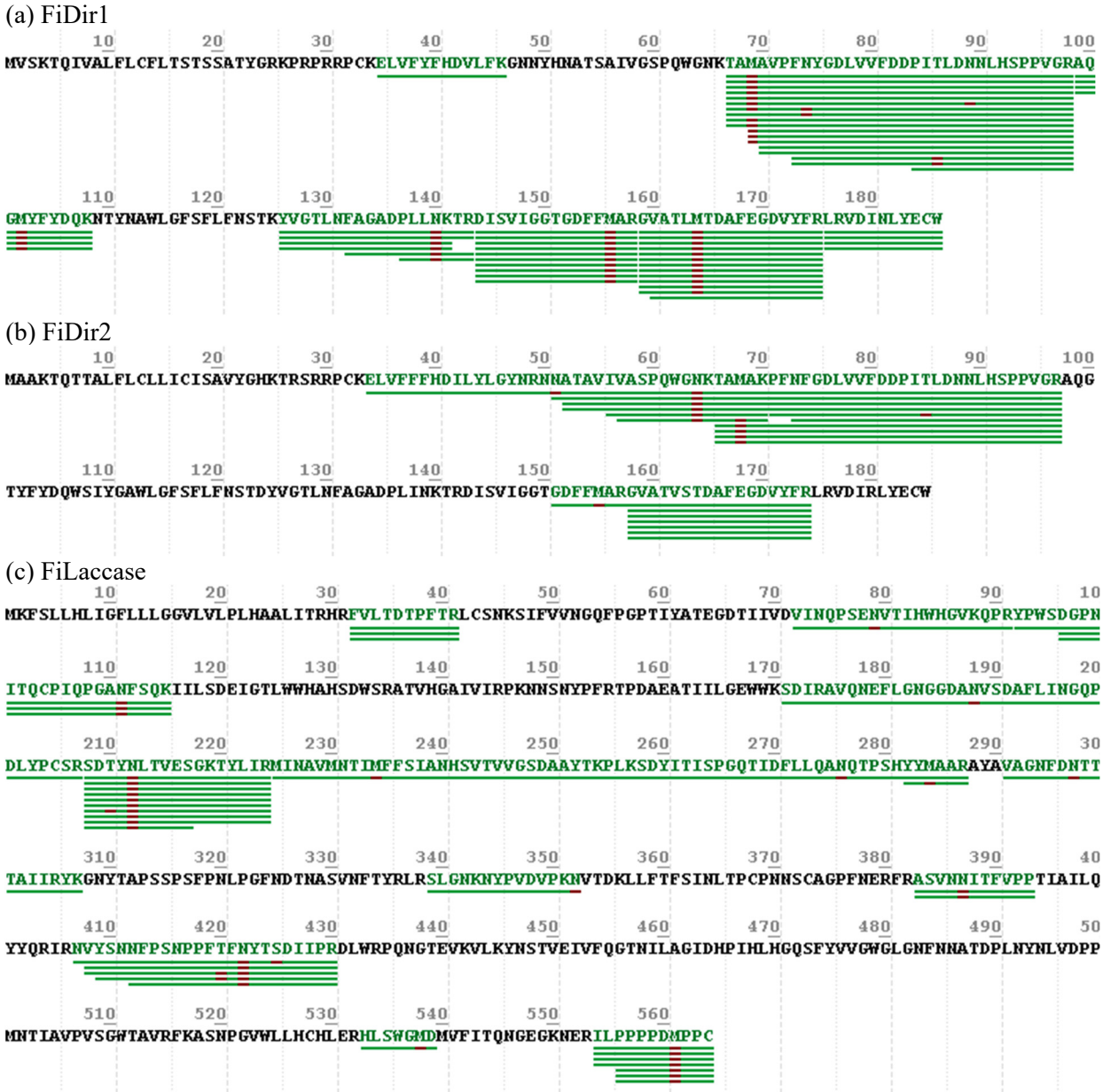
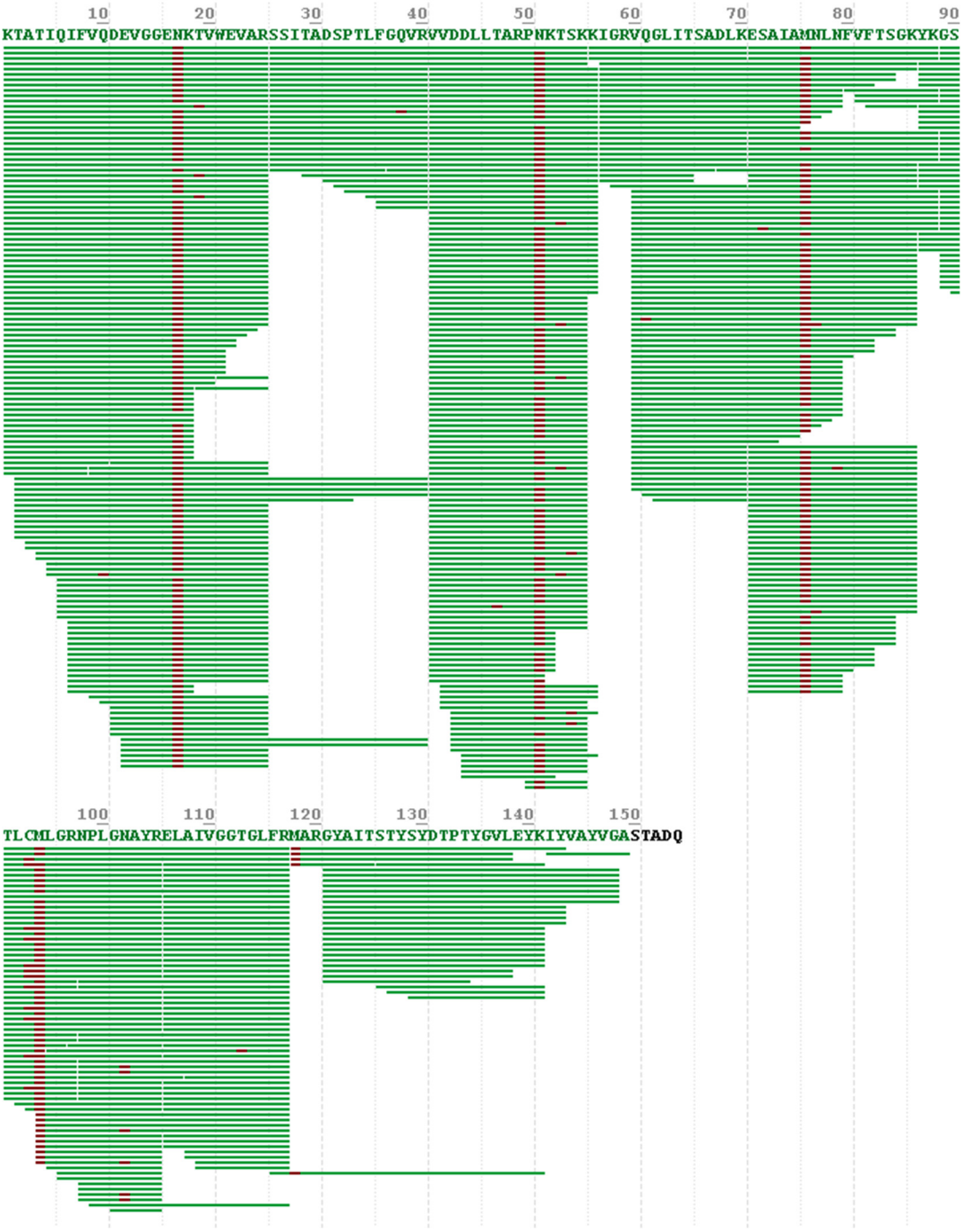


Figure S5. (a) FiDir1, (b) FiDir2, and (c) FiLaccase sequence coverage based on peptide mapping result in Byonic. Green bars indicate peptides are detected in the sequence window. Pink regions indicate modifications are detected at the residue. Note that some peptides from FiDir1 and FiDir2 are identical (due to high sequence identity) and are only assigned to FiDir1. Glycosylation was detected at many of the N residues. The sequence coverages were low for all of them, indicating their low abundance in the sample. None of them were observed above detection limit in native MS or top-down. Highly heterogenous glycosylation may be present, or they could be binding to heterogeneous proteins. Heterogeneity would significantly "dilute" the signal at the intact protein level, making them hard to detect.



Major N17 glycosylation: HexNAc(2)Hex(3)Fuc(1)Pent(1) 1170.42 Da; HexNAc(2)Hex(3)Fuc(1) 1038.38 Da; HexNAc(2)Hex(3)Pent(1) 1024.36 Da.

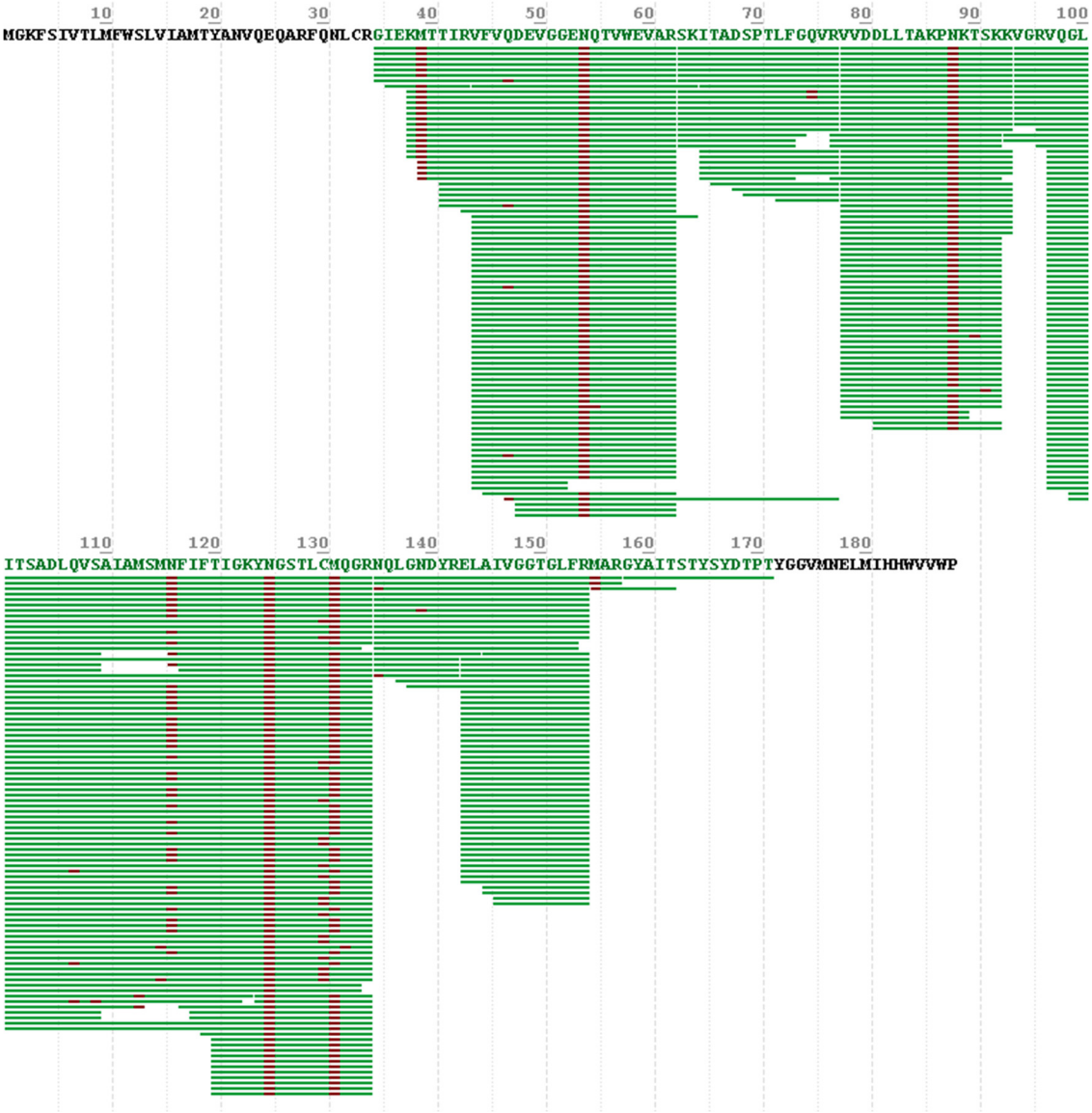
Major N51 glycosylation: HexNAc(2)Hex(3)Fuc(1)Pent(1) 1170.42 Da; HexNAc(4)Hex(3)Pent(1) 1430.52 Da; HexNAc(2)Hex(3)Fuc(1) 1038.38 Da; HexNAc(3)Hex(4)Fuc(1)Pent(1) 1535.55 Da.

Figure S6. FiDir18 sequence coverage based on peptide mapping result in Byonic. The C-terminus is not covered completely.

(a) Peptide coverage for FiDir19 based on best hit to *Forsythia koreana* transcript.

Note: top down data showed the main proteoform has the starting residue K37 in this sequence.

(b) Peptide coverage for FiDir19 based on *Forsythia suspensa* genome.



Note: top down data showed the main proteoform has the starting residue K38 in this sequence.

Major glycosylation at N54 (or N17 in actual protein): HexNAc(2)Hex(3)Fuc(1)Pent(1) 1170.42 Da; HexNAc(2)Hex(3)Pent(1) 1024.36 Da; HexNAc(2)Hex(3)Fuc(1) 1038.38 Da; HexNAc(2)Hex(3) 892.32 Da.

Major glycosylation at N88 (or N51 in actual protein): HexNAc(2)Hex(3)Fuc(1)Pent(1) 1170.42 Da; HexNAc(2)Hex(3)Fuc(1) 1038.38 Da; HexNAc(2)Hex(3)Pent(1) 1024.36 Da.

Major glycosylation at N125 (or N88 in actual protein): HexNAc(2)Hex(3)Fuc(1)Pent(1) 1170.42 Da.

Figure S7. FiDir19 sequence coverage to sequence predicted from (a) *Forsythia koreana* transcript, and (b) *Forythia suspensa* genome. They only differ at the C-terminal residues, as they do not have experimental coverage to confirm. Based on top-down data, the (b) sequence is more likely.

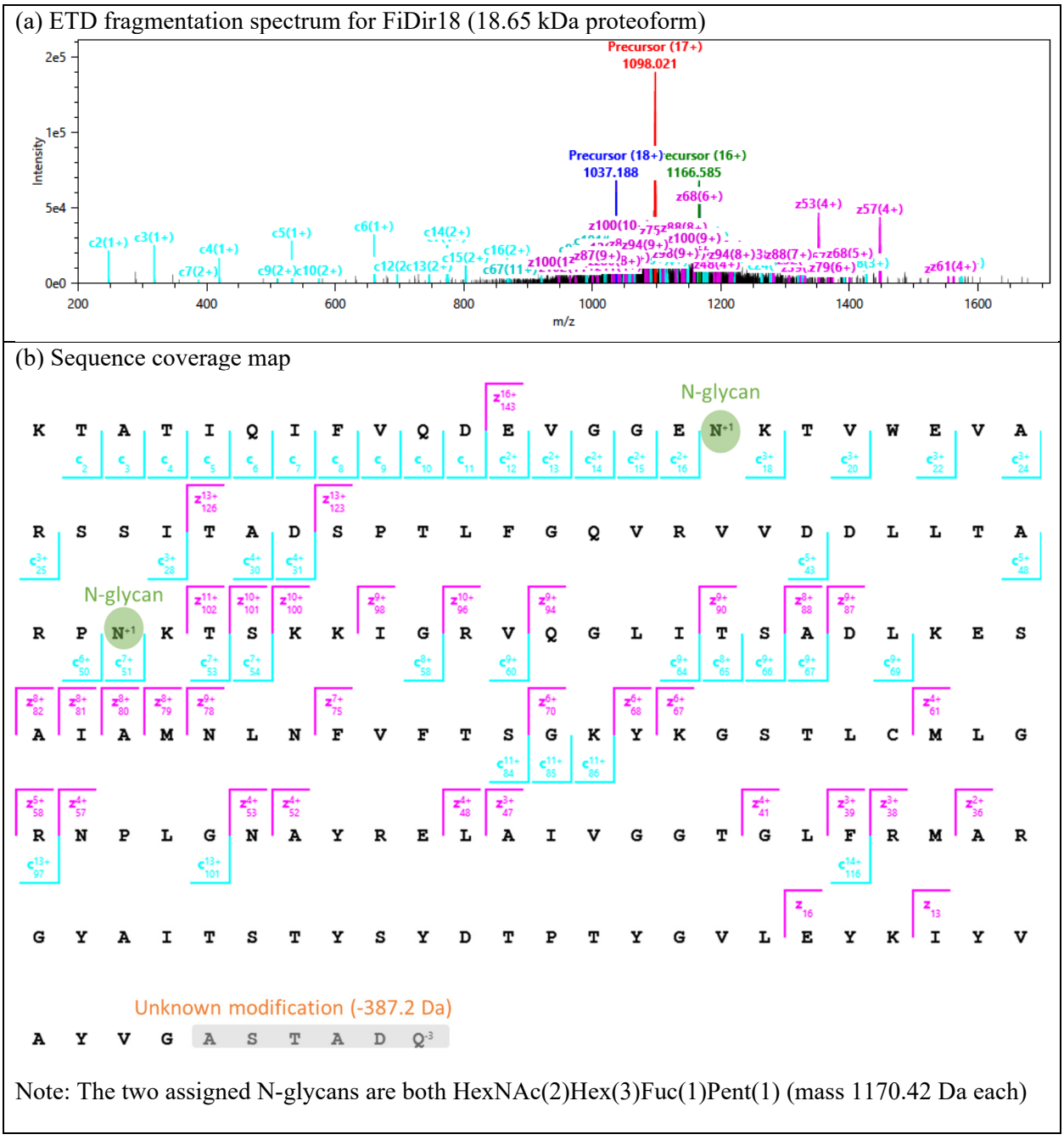


Figure S8. (a) FiDir18 top-down fragmentation spectrum and (b) sequence coverage map. Monoisotopic mass of the proteoform is 18637.2 Da. Cyan wedges represent c-ion coverage from the N-terminus, and pink wedges represent z-ion coverage from the C-terminus. Subscripts show the c/z ion number, superscripts show the charge states of the matched fragment ions. The coverage map was generated with LcMsSpectator, and the same format was used for other top-down coverage maps in this manuscript.

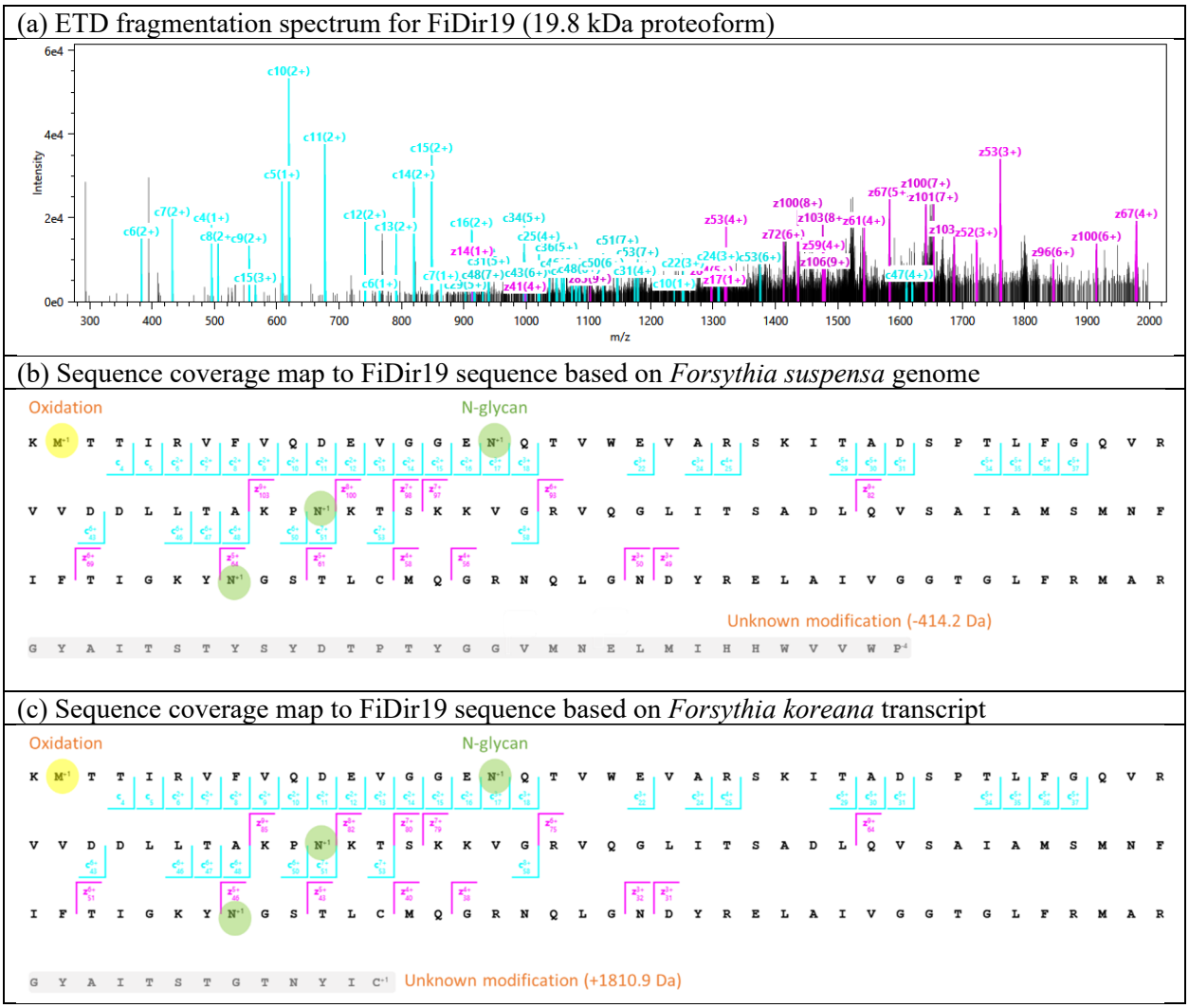


Figure S9. (a) FiDir19 top-down fragmentation spectrum and (b) sequence coverage map matching to the sequence predicted from *Forsythia suspensa*. The sequence coverage map matching to the sequence predicted from *Forsythia koreana* transcript is in (c). Modifications are highlighted in (b-c) for oxidation and N-glycosylation. All three assigned N-glycans are HexNAc(2)Hex(3)Fuc(1)Pent(1) (mass 1170.42 Da each). Note that the sequence in (c) is shorter than (b), thus the numbering of z ions is different. Most protein sequences can be confirmed. However, the C-termini cannot be matched by either top-down or peptide data and may contain unknown modifications. The other possibility is that DP sequences in *Forsythia* × *intermedia* are significantly different from other *Forsythia* species, at least in the C-terminal region. Given the larger unexplainable mass in (c), the sequence in (b) is more likely.

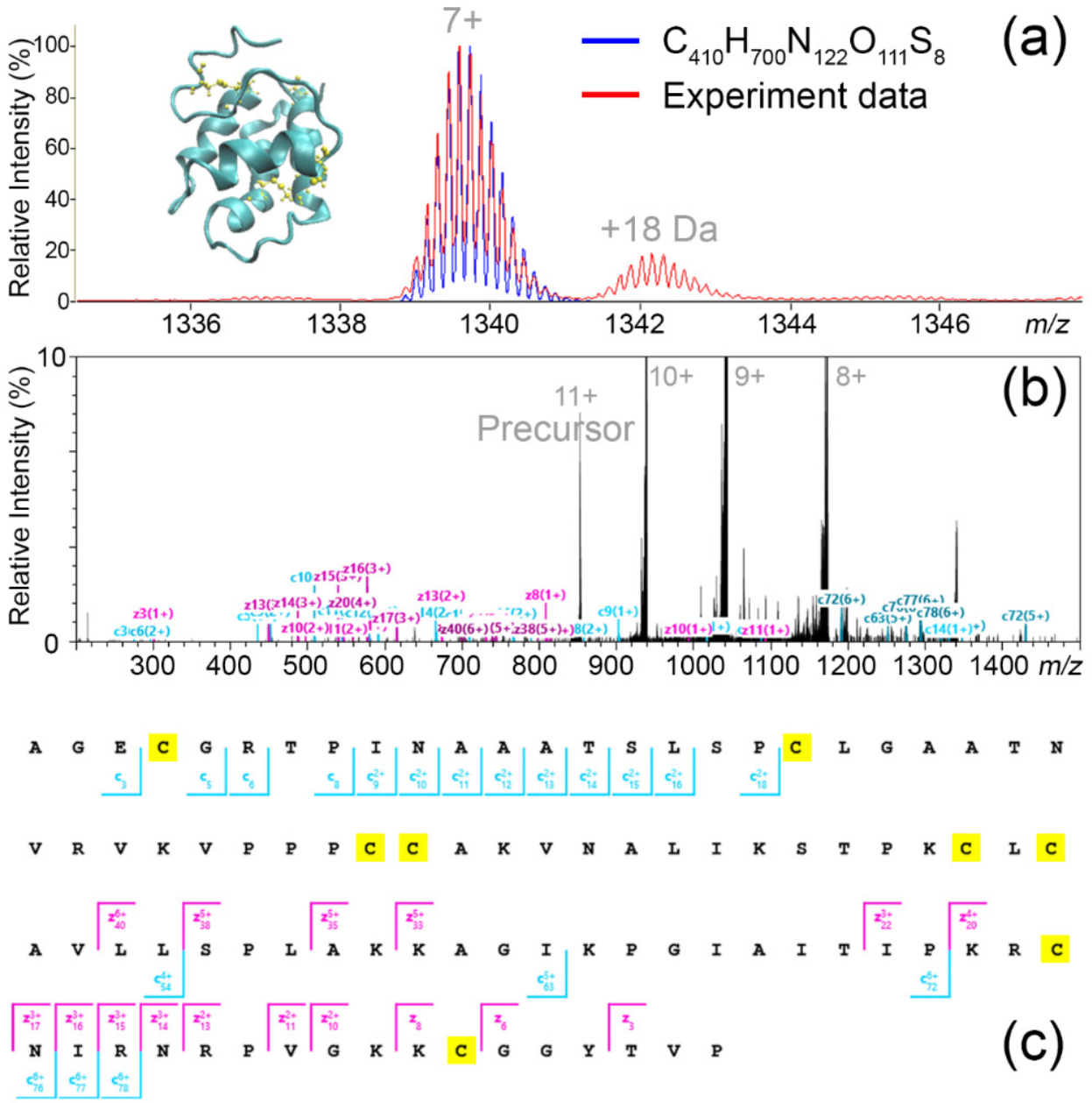


Figure S10. (a) Intact mass spectrum of nsLTP. Red line is the experimental data, overlaid with the theoretical isotope distribution in blue. The inset structure is the homology model built in I-TASSER, with cysteine residues highlighted in yellow. (b) ETD fragmentation spectrum of intact nsLTP. The disulfide bonds reduced the yield of fragments. The charge reduced precursors were the dominant peaks. To better display the fragments, the y axis was only shown up to 10%. (c) Sequence coverage was based on the ETD fragments. Cysteine residues are highlighted in yellow. The 9.4 kDa nsLTP was consistently detected as one of the major components in the extract. The intact mass suggests there was an –8 Da shift from the unmodified sequence mass, meaning all 8 cysteines were disulfide bonded. We also observed partially reduced forms of nsLTP in the intact protein LCMS data (not shown). Because we did not intentionally preserve the oxidation state in its sample preparation, we do not yet know what the active forms are *in vivo*. Despite the extensive disulfide bonding, top-down fragmentation using ETD still yielded sufficient ions to confirm the sequence at the intact protein level. nsLTP was also confirmed in bottom-up data (not shown).

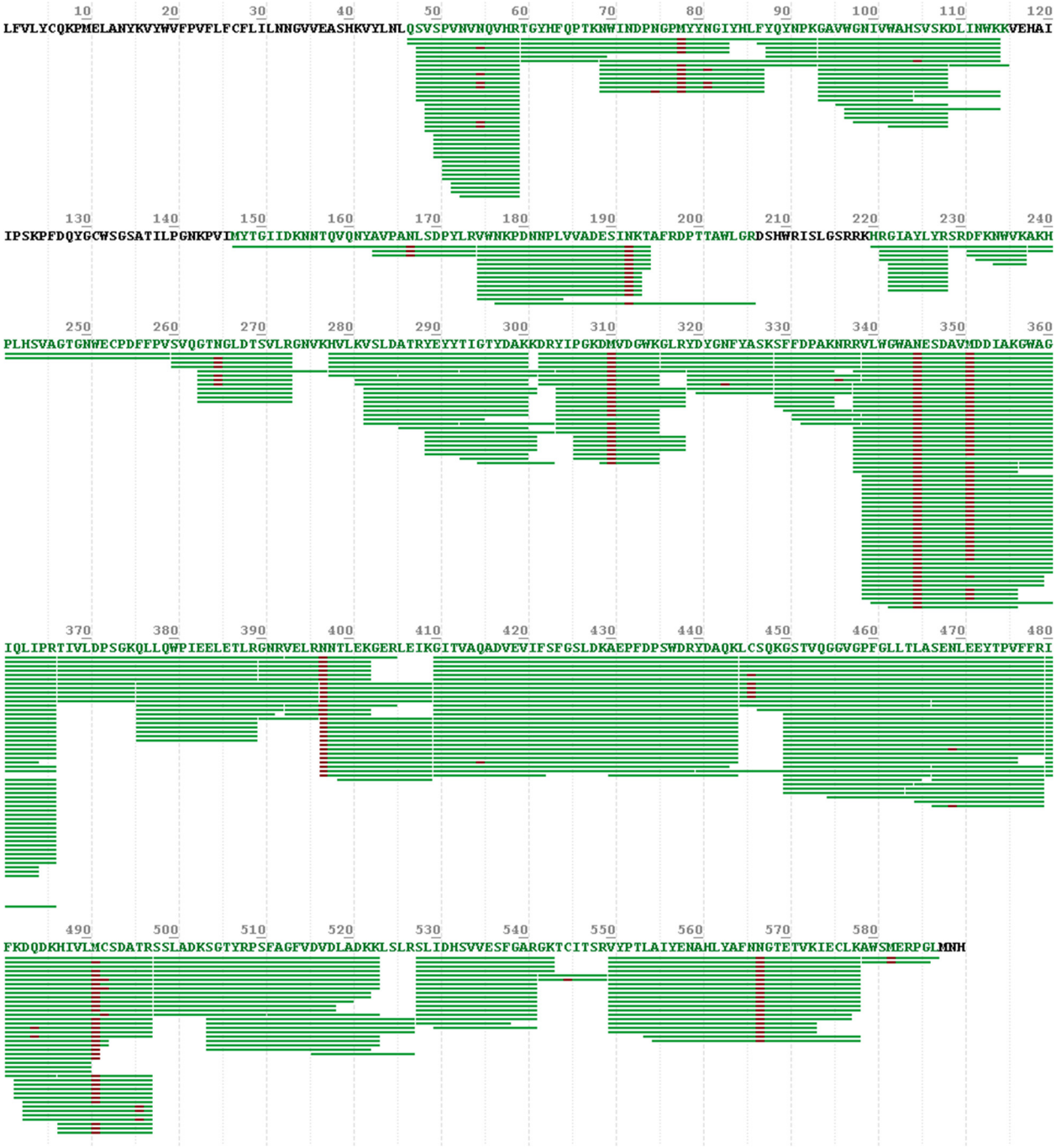


Figure S11. Beta-fructofuranosidase (invertase) sequence coverage based on peptide mapping result in Byonic. Glycosylation is detected at N192, N345, N397, N567.

Major glycans near N102: HexNAc(2)Hex(3)Fuc(1)Pent(1) 1170.42 Da; HexNAc(2)Hex(3)Pent(1) 1024.36 Da; HexNAc(2)Hex(2) 730.26 Da; HexNAc(2)Hex(1) 568.21 Da; HexNAc(2) 406.16 Da.

Major glycans near N217: HexNAc(2)Hex(3)Fuc(1)Pent(1) 1170.42 Da; HexNAc(2)Hex(2) 730.26 Da; HexNAc(2)Hex(1) 568.21 Da; HexNAc(1) 203.08 Da.

Figure S12. Peroxidase sequence coverage based on peptide mapping result in Byonic. Glycosylation was detected near N102 and N217. The current data cannot confidently assign the glycosylation site to differentiate N-glycans and O-glycans. Other modifications are methionine oxidation and deamidation. The actual protein sequence likely starts at A31 based on the coverage. The average mass of the most probable amino acid sequence (Table S1) is 32193 Da, ~97 Da less than the major species with the lowest mass in Figure 2c. The 97 Da may be due to PTM or salt addition (e.g. sulfate, used in the purification buffer). Therefore the 32.29 kDa species may be without glycosylation and/or with truncations. Other higher mass species are different glycosylated forms.

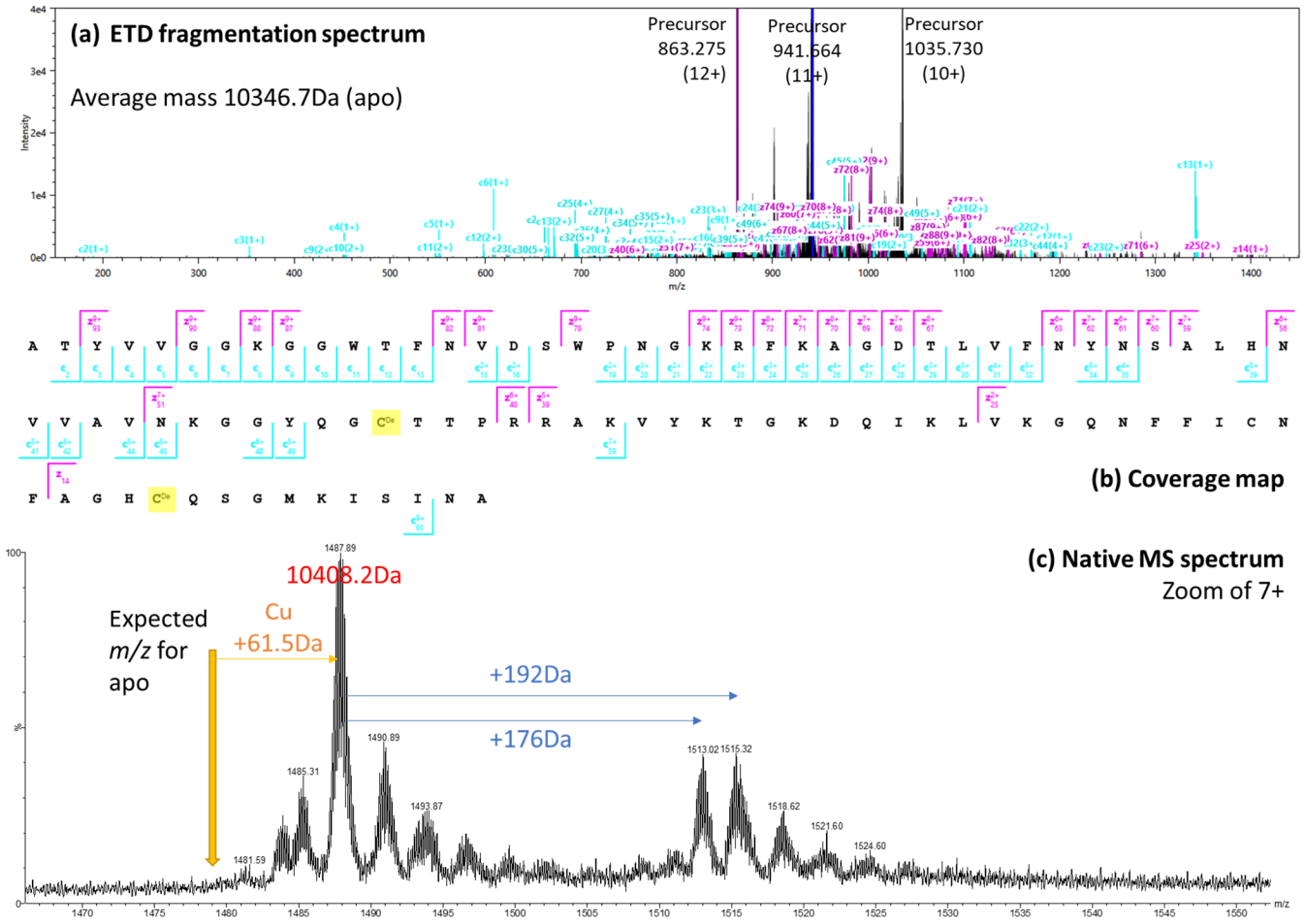
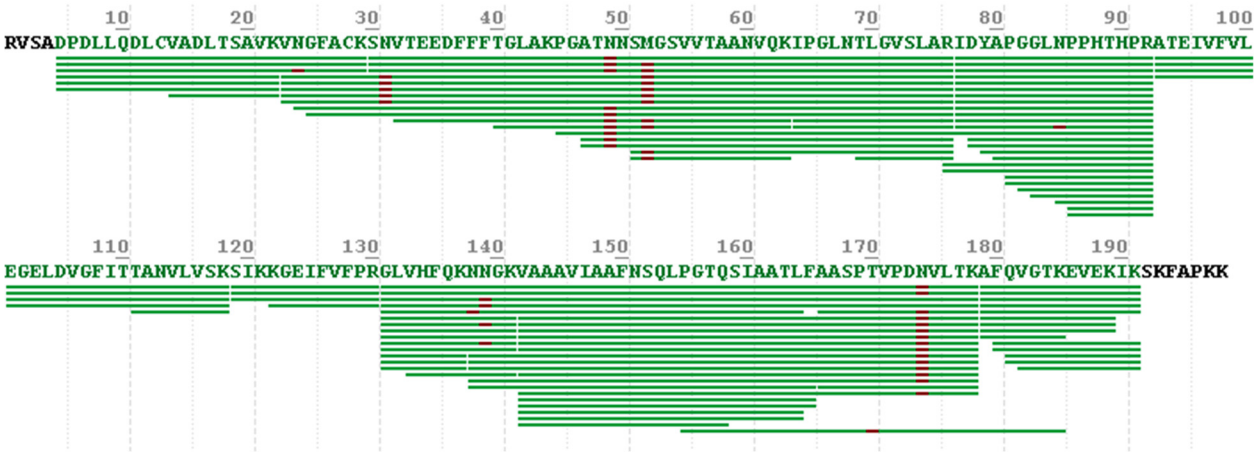


Figure S13. (a) Top-down fragmentation spectrum for denatured cupredoxin family protein. (b) Sequence coverage map for the top-down data, confirming most of the residues. Disulfide was applied on the two cysteines highlighted in yellow (–2 Da). (c) Native MS spectrum of the same protein. The *m/z* of the denatured protein is expected at 1479 as shown by the yellow arrow. However, the major peak is shifted by 61.5 Da, which matches one bound Cu^{2+} (mass of Cu minus two protons). Other peaks correspond to mass shifts with salt adducts, oxidation, water loss. The +192 Da and +176 Da peaks may be the same protein with unknown modifications.



Major glycans near N31: HexNAc(3)Hex(3)Fuc(1)Pent(1) 1373.50 Da.

Figure S14. Germin-like protein 1 sequence coverage based on peptide mapping result in Byonic.

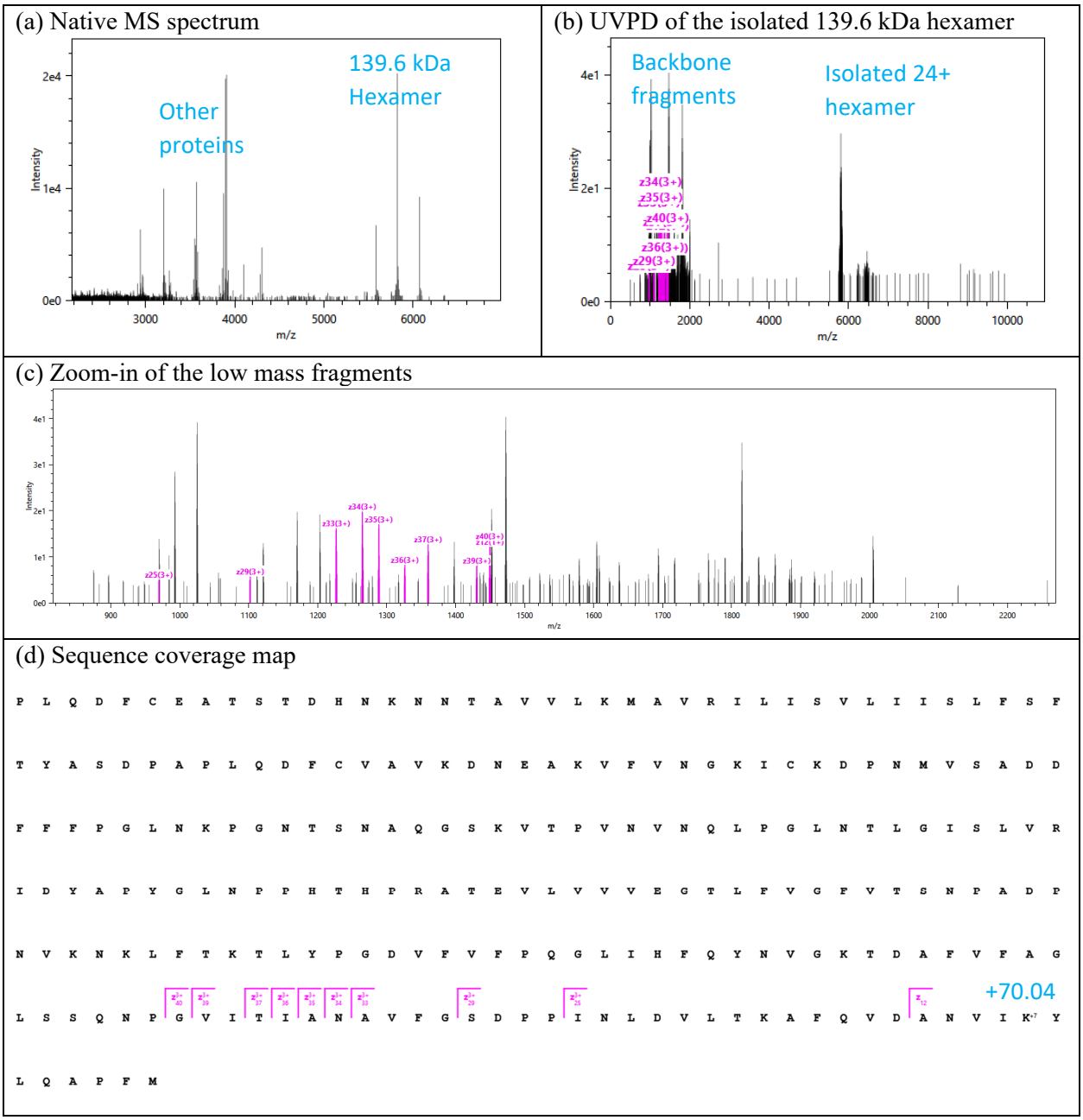


Figure S15. Native top down analysis indicated the hexamer is likely a germin-like protein 2. The same sample in Figure S4 (F1 high mass region showed ~140k Da species) was analyzed on a modified Thermo UHMR Orbitrap with the same nano-electrospray conditions.[3,4] The heated capillary in the source was at 250 °C. In source trapping of 100 V was used to bias the detection of the 139.6 kDa hexamer over the smaller protein species. The 139.6 kDa hexamer was mass isolated, and subjected to 157 nm UVPD (2 mJ 5 pulses). Despite low coverage, the detected fragments can be matched to the C-terminal residues of the sequences predicted from *de novo* sequencing. A mass shift of 70.04 Da was assigned to the C-terminus to match the fragments. The 70.04 Da was tentatively assigned to lysine crotonylation, but can be other combinations of modifications of amino acid substitutions. C-termini of germin is more exposed than N-termini in known structures (e.g. PDB 1FI2). Therefore, preferential cleavage of C-terminal residues was expected in native top down based on the surface accessibility.

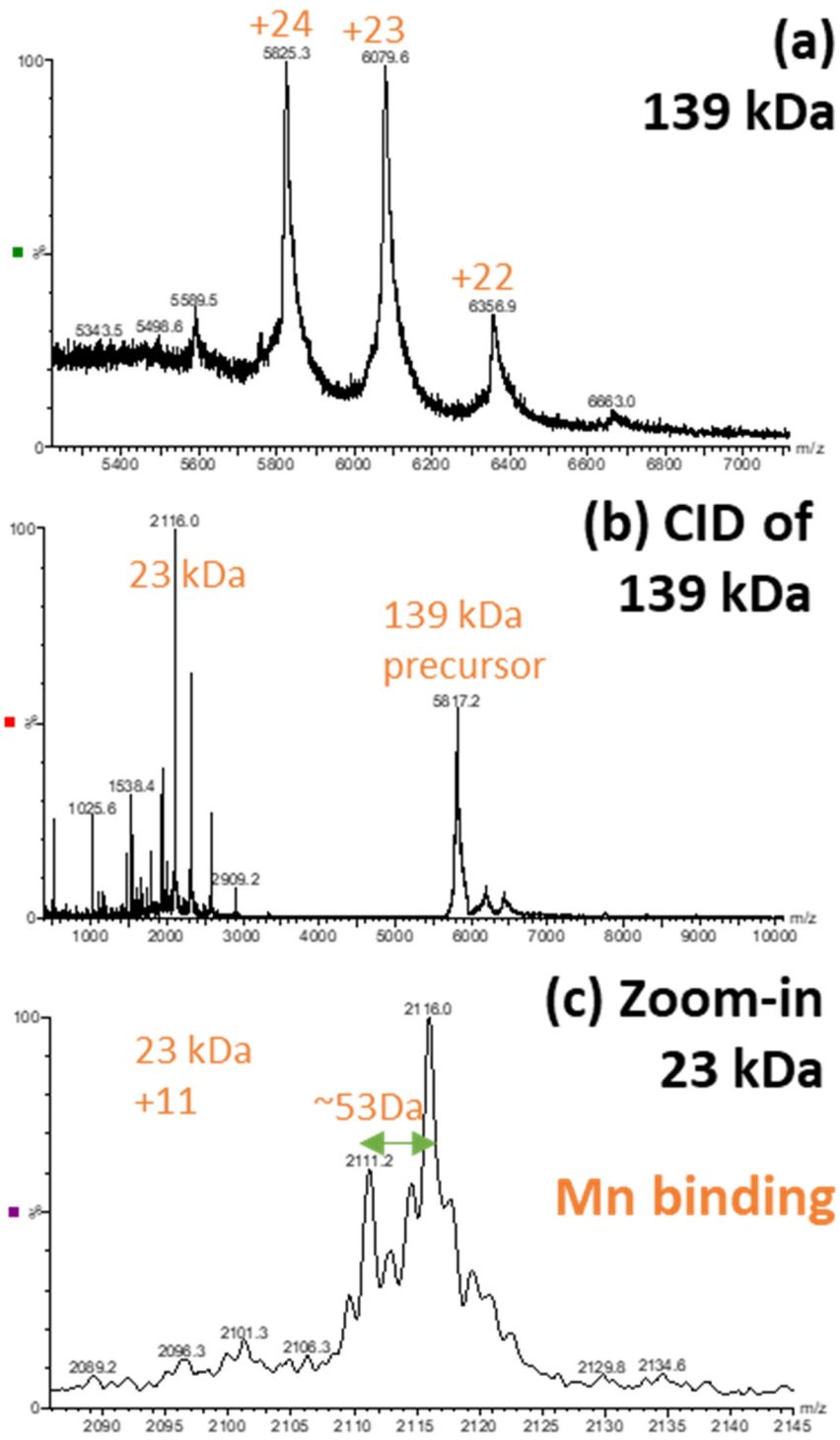


Figure S16. Native MS data for the germin-like protein 2 hexamer: (a) intact 139 kDa complex, data extracted from Figure S4 F1. (b) CID of the isolated 139 kDa species, releasing 23 kDa monomer. (c) Zoom in of the 23 kDa monomer showed mass differences of ~53 Da between apo and Mn binding peaks. The assignment to Mn is based on the mass (~55 Da minus 2 protons for Mn^{2+}) and the known substrate for germin.

(a) Precursor - **FiDir18: FiDir19 2: 1**
57.1 kDa, 14+, *m/z* 4080

	Mass (Da)	Intensity	Area
○	18,490	9.63	17.21
▽	18,650	100.00	100.00
△	18,810	5.29	8.12
▷	19,800	11.79	17.45
▢	19,940	7.78	13.90

(b) Precursor - **FiDir18: FiDir19 1: 2**
58.3 kDa, 14+, *m/z* 4160

	Mass (Da)	Intensity	Area
○	18,490	15.67	26.92
▽	18,650	100.00	100.00
△	18,810	8.16	17.87
▷	19,650	12.87	24.03
▢	19,800	62.67	75.72
◇	19,960	19.55	30.06

(c) Precursor - **FiDir18: FiDir19 0: 3**
59.3 kDa, 14+, *m/z* 4234

	Mass (Da)	Intensity	Area
○	18,650	14.41	17.36
▽	19,650	29.36	39.45
△	19,800	100.00	100.00

(d) Precursor - **FiDir18: FiDir19 3: 0**
56.0 kDa, 13+, *m/z* 4300

	Mass (Da)	Intensity	Area
○	18,650	100.00	100.00

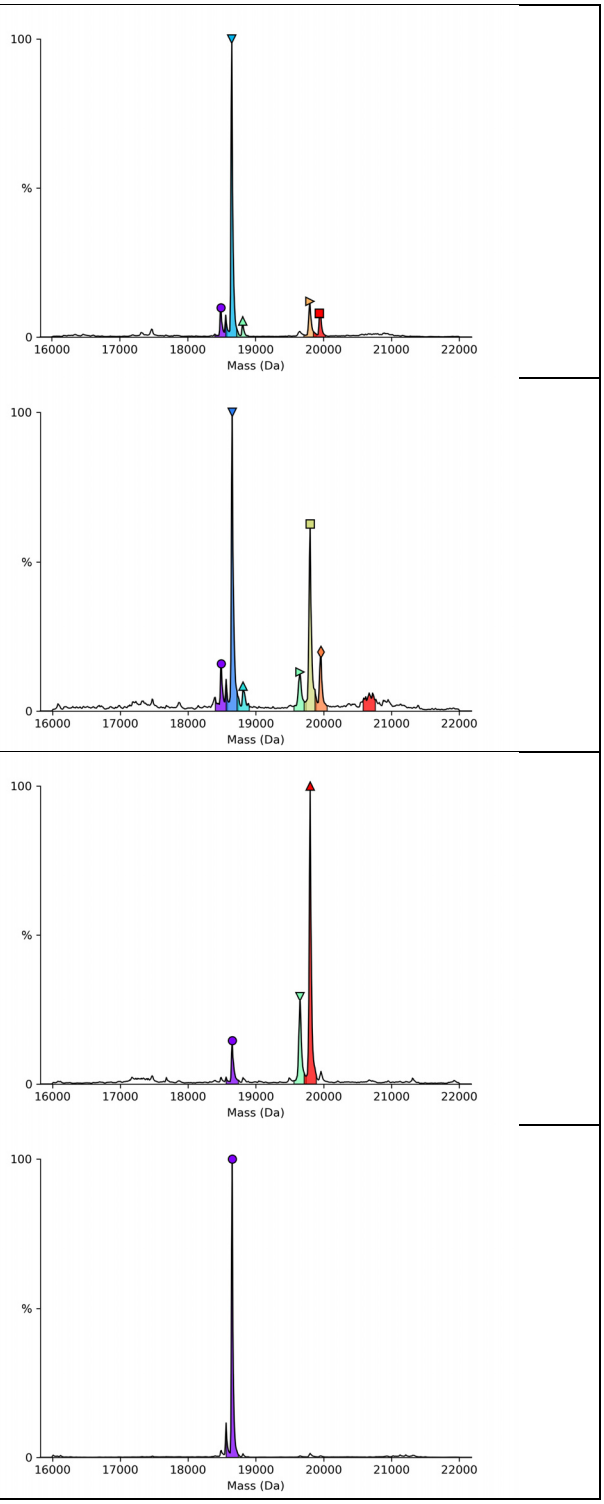


Figure S17. Released DP monomers from CID of different FiDir18/FiDir19 trimers. The left column shows information on the precursor and the list of the released monomers (symbol, mass, relative intensity/area). The right column shows deconvoluted mass spectra. Peaks are annotated with symbols that match the table in the left column in each row. The ratio of released FiDir18:FiDir19 monomers scales with the precursor stoichiometry. It is noted that the homo-trimer of FiDir19 in (c) showed some FiDir18 monomer because of co-isolation of other species in the same *m/z* window.

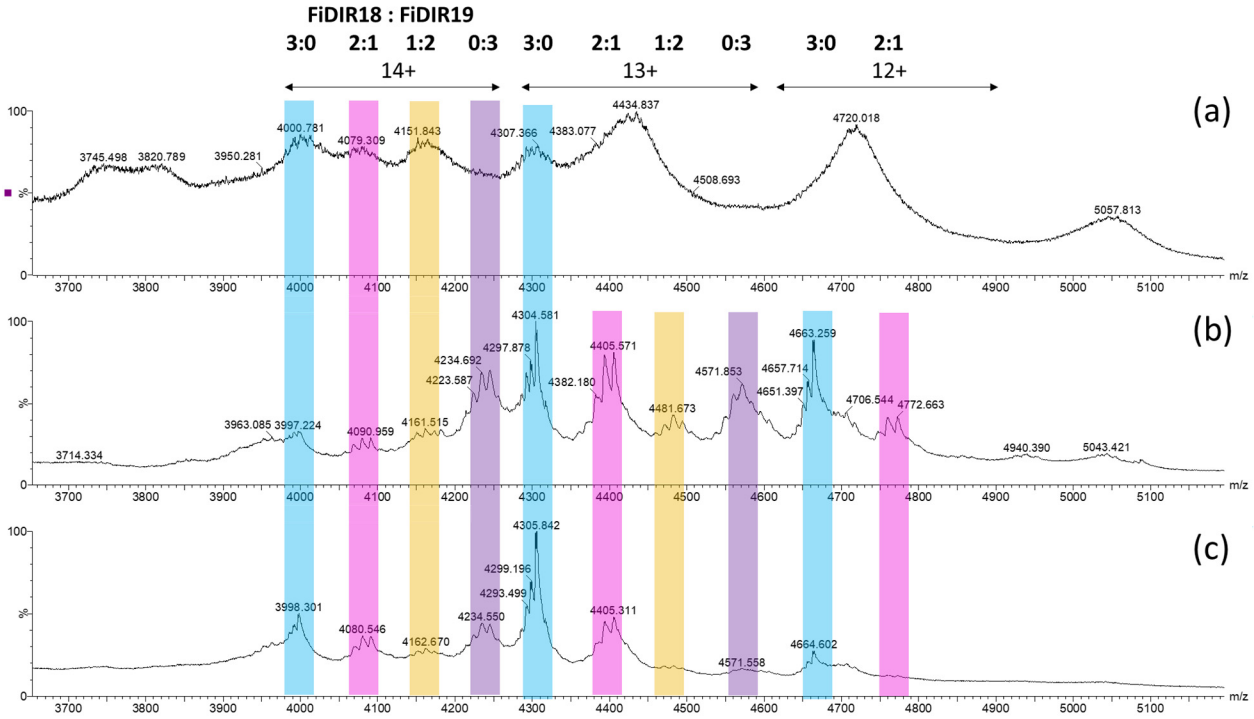


Figure S18. Native MS spectra of 3 biological replicates all detected hetero-trimers between FiDir18 and FiDir19 as discussed in Figure 4. Peaks assigned to DP trimers are highlighted in colored bars. Each color represents the same species at different charge states. Peak identity and charge states are labeled at the top. Data for replicate (c) is the same as Figure 4. Replicate (a) is from the initial MonoS fraction, and other protein species overlapped significantly with the DP peaks. In addition, sample (a) was more "salty", leading to more nonspecific adducts and broader peaks than the secondary fractions in (b-c). However, the major species detected were qualitatively the same.

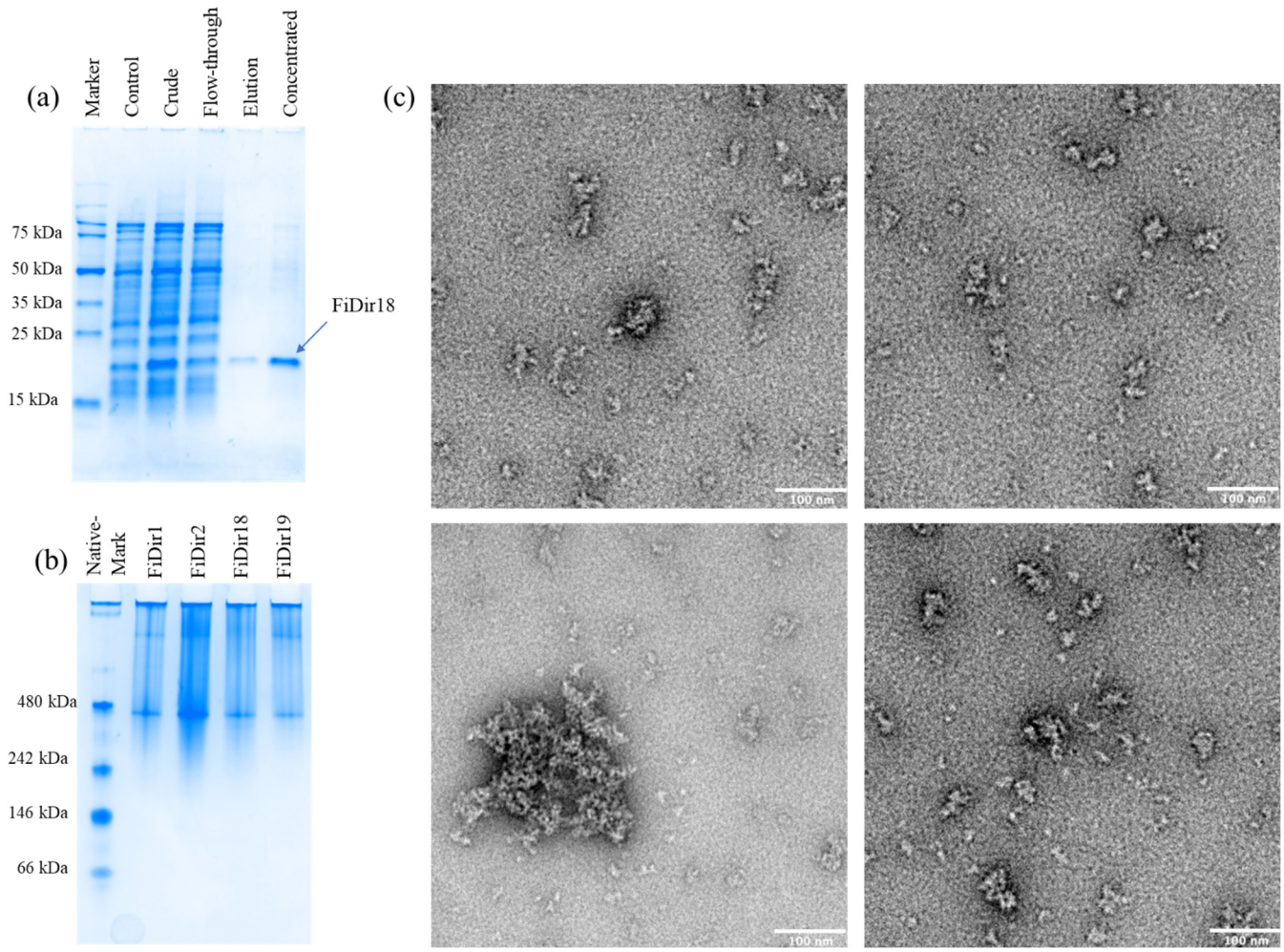


Figure S19. Cell-free protein expression of dirigent proteins and their characterization. (a) A representative 10% SDS-PAGE gel for the synthesis and 3XFLAG-purification of FiDir18 homolog, stained with Bio-Safe Coomassie Blue (BIO-RAD). Control lane represents a parallel experiment conducted where DNA template was not supplemented. SDS-PAGE gels for other homologs look similar and are not shown here. (b) Native PAGE (4-15% gradient) of all synthesized proteins. (c) Transmission electron images collected on ETEM (FEI) for FiDir18. The protein was deposited at 50 µg/ml concentration on the glow-discharged ultrathin carbon grid (Ted Pella, Inc) and stained for 1 min with NanoW (Nanoprobes). TEM images for other DP homologs were also collected and showed the same aggregation (data not shown).

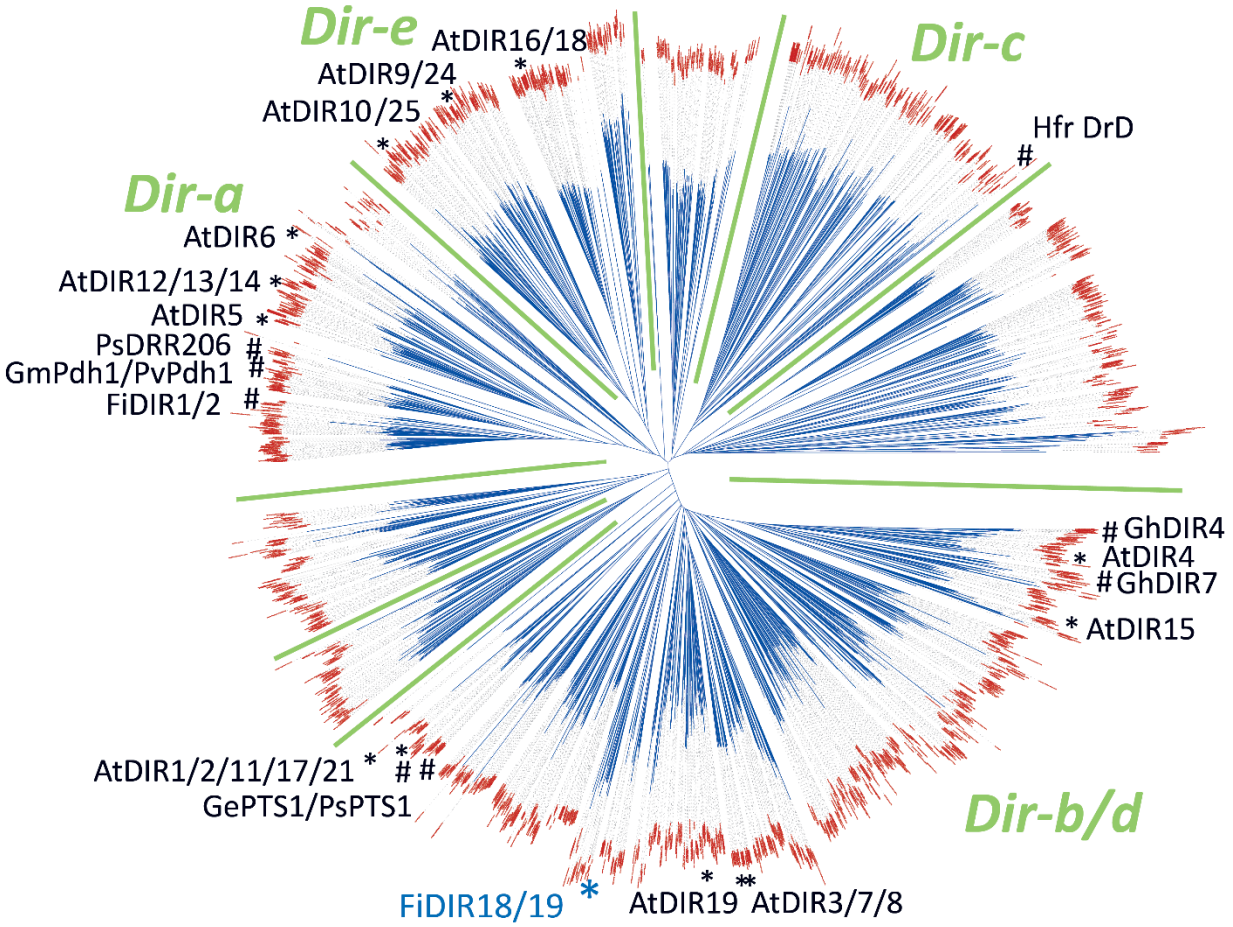


Figure S20. Unrooted phylogenetic tree of the dirigent/dirigent-like protein superfamily constructed from several thousand sequences. Major subfamily divisions are indicated with green lines, and the Dir-a, Dir-b/d, Dir-c, and Dir-e are identified by name. Key landmark sequences are indicated with asterisks (for *Arabidopsis* DPs) and hash symbols (for other characterized DPs). More details of the characterized DPs are included in Table S3. The new *Forsythia* dirigent proteins identified in this work are indicated in blue. The tree was rendered with iTOL[5] using a sequence alignment and tree generated with Clustal Omega.[6] FiDir18 and FiDir19 are sufficiently distant from most other Dir-b/d sequences that it may be problematic to place them accurately in the overall DP phylogenetic tree.

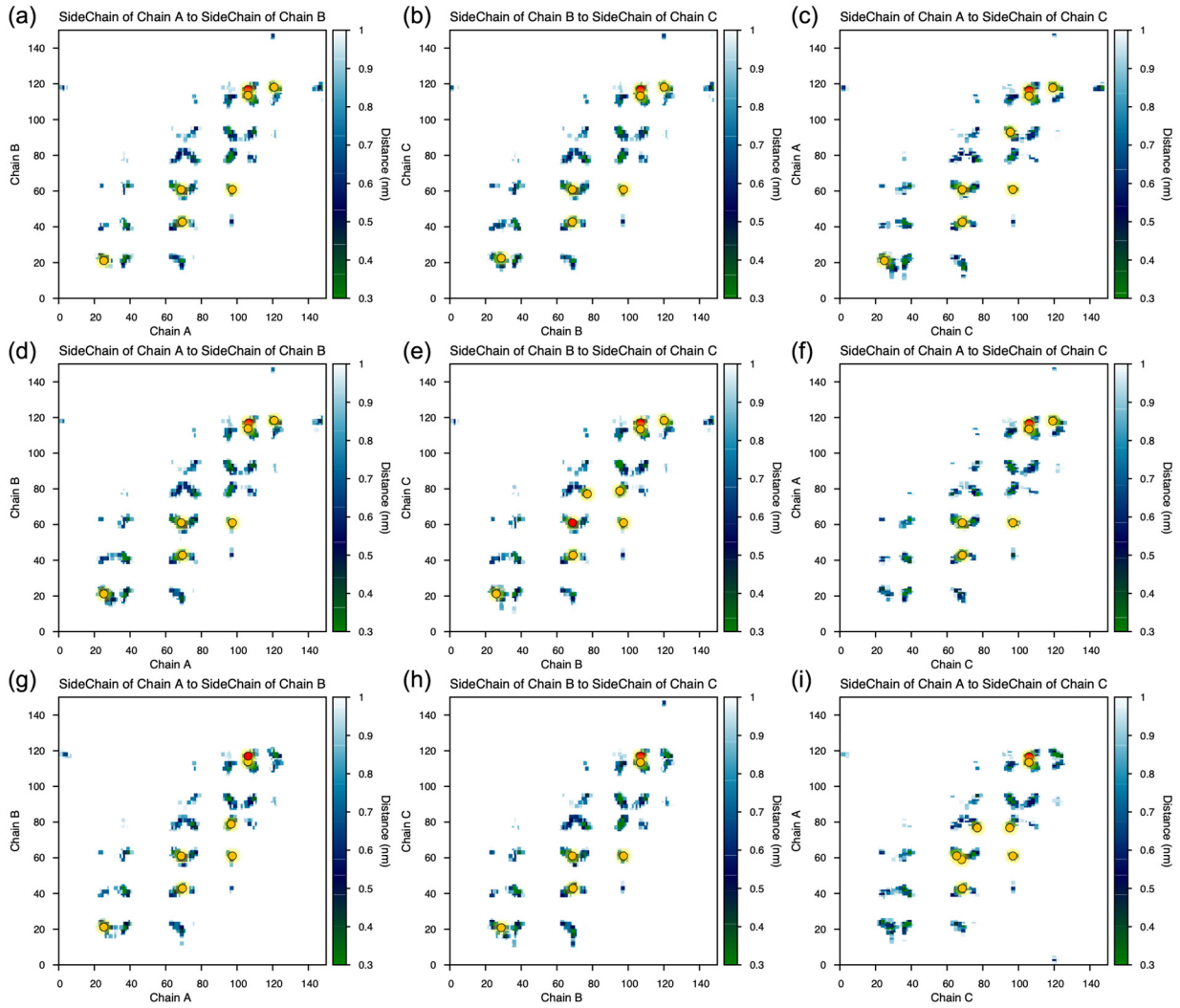


Figure S21. Side-chain to side-chain interactions for the FiDir18 homo-trimeric system (top row), and FiDir18/19 hetero-trimeric systems (middle row = FiDir18 two monomers A/B, FiDir19 one monomer C; bottom row = FiDir19 two monomers A/B, FiDir18 one monomer C) calculated from the MD simulations. The FiDir18 system is shown in panels (a-c) where the interactions between Chain A to chain B are shown in (a), chain B to chain C are shown in (b), and chain C to chain A are shown in (c). The FiDir18 (two monomers A/B) /FiDir19 (one monomer C) system is shown in (d-f) where the interactions between Chain A to chain B are shown in (d), chain B to chain C are shown in (e), and chain C to chain A are shown in (f). The FiDir19 (two monomers A/B) /FiDir18 (one monomer C) system is shown in panels (g-i) where the interactions between Chain A to chain B are shown in (g), chain B to chain C are shown in (h), and chain C to chain A are shown in (i). Values for the distances between the interactions are used to color the heat map, with 3.0 Å colored as green and > 10 Å colored as white. Interactions beyond 10 Å are not shown. Hydrogen bonding interactions are shown as yellow circles and indicate a hydrogen bond with an occupancy >20%. Red circles denote side chains involving a salt bridge interaction.

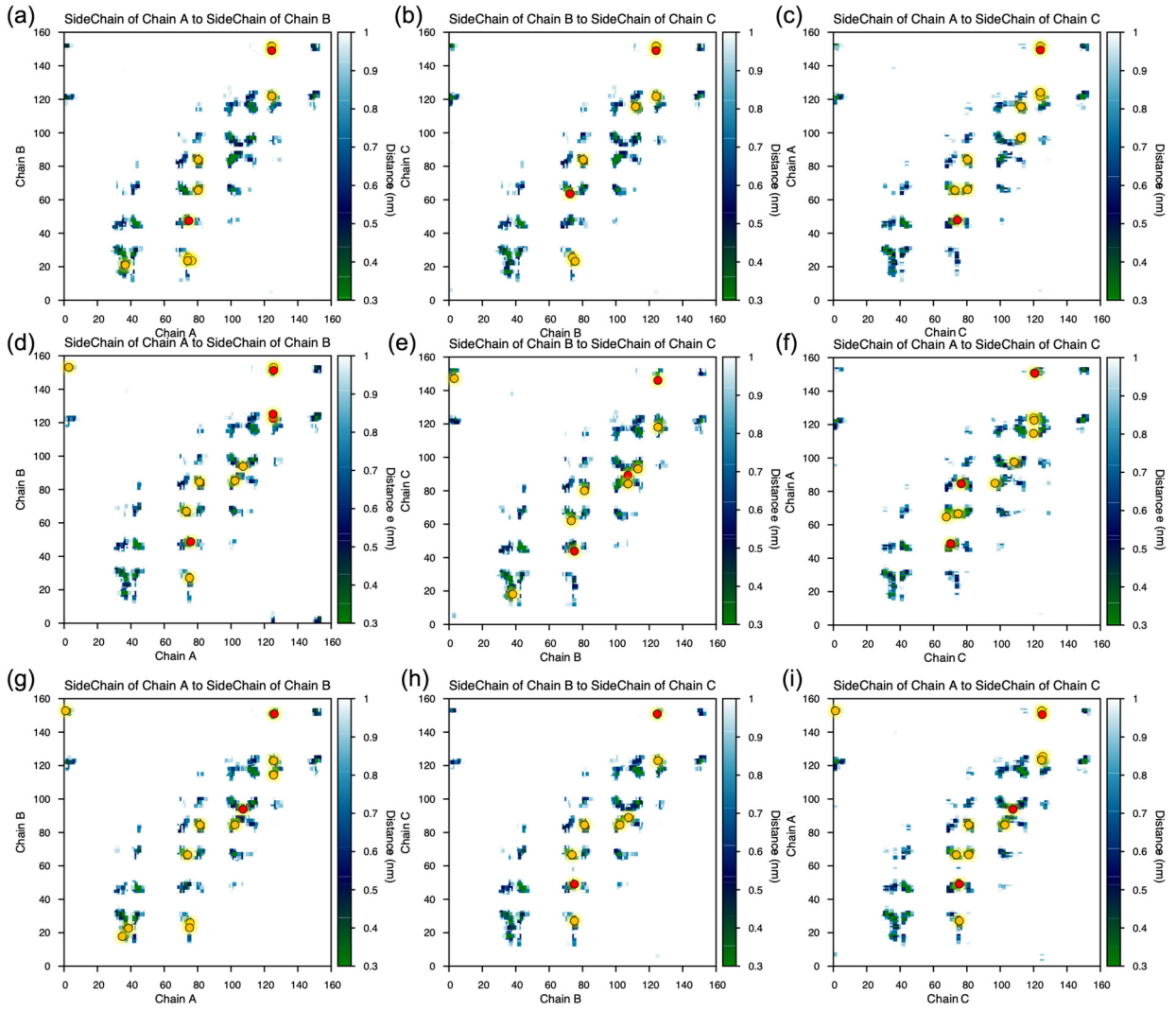


Figure S22. Side-chain to side-chain interactions for homo-trimeric systems for AtDir5 (top row) and AtDir6 (bottom row). Hetero-trimeric system AtDir6/AtDir5 consisting of two monomers of AtDir6, chains A and B, and one monomer of AtDir5, chain C (middle row). The AtDir5 system is shown in (a-c) where the interactions between Chain A to chain B are shown in (a), chain B to chain C are shown in (b), and chain C to chain A are shown in (c). The AtDir6 (two monomers A/B) /AtDir5 (one monomer C) system is shown in (d-f) where the interactions between Chain A to chain B are shown in (d), chain B to chain C are shown in (e), and chain C to chain A are shown in (f). The AtDir6 homo-trimeric system is shown in (g-i) where the interactions between Chain A to chain B are shown in (g), chain B to chain C are shown in (h), and chain C to chain A are shown in (i). values for the distances between the interactions are used to color the heat map, with 3.0 Å colored as green and > 10 Å colored as white. Interactions beyond 10 Å are not shown. Hydrogen bonding interactions are shown as yellow circles and indicate a hydrogen bond with an occupancy >20%. Red circles denote side chains involving a salt bridge interaction. For clarity, each protein sequence used in this analysis was renumbered to begin at 1.

References

1 J. J. Almagro Armenteros, K. D. Tsirigos, C. K. Sønderby, T. N. Petersen, O. Winther, S. Brunak, G. von Heijne and H. Nielsen, *Nat. Biotechnol.*, 2019, **37**, 420–423.

2 L.-F. Li, S. A. Cushman, Y.-X. He and Y. Li, *Hortic. Res.*, 2020, **7**, 130.

3 J. B. Shaw, W. Liu, Y. V. Vasil′ev, C. C. Bracken, N. Malhan, A. Guthals, J. S. Beckman and V. G. Voinov, *Anal. Chem.*, 2019, **92**, 766–773.

4 M. Zhou, W. Liu, J. B. J. B. Shaw and J. B. Shaw, *Anal. Chem.*, 2020, **92**, 1788–1795.

5 I. Letunic and P. Bork, *Nucleic Acids Res.*, 2021, **49**, W293–W296.

6 F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson and D. G. Higgins, *Mol. Syst. Biol.*, 2011, **7**, 539.